Differentially Private Ranking Release for Kernel SHAP: A Certified Exponential-Mechanism Approach with Empirical Sensitivity Diagnostics

Bader Alissaei*
VaultBytes Innovations Ltd.

Abstract

We study input-level (background-record) differentially private release of kernel SHAP explanations. The lead positive result is a *ranking release mechanism* (§4) whose privacy is pure ε-DP under any valid upper bound on the per-coordinate ranking sensitivity Δ_∞, and whose utility is governed empirically by the dimensionless ratio Δ_∞/g between Δ_∞ and the top-1/top-2 magnitude gap g.

Lead positive result. The *exponential mechanism* for top-1 release, $\Pr[i^* = i] \propto \exp(\varepsilon|\varphi_i|/(2\Delta_\infty))$, is pure ε-DP [Dwork and Roth, 2014] and satisfies

$$\Pr[i^* \text{ correct}] \;\geq\; 1 - d\exp\left(-\frac{\varepsilon g}{2\Delta_\infty}\right).$$

Closed-form Δ_∞ certificates are given for linear and logistic models; for general models we use a conservative analytic bound or a non-certified empirical diagnostic (§4, §4.2). *Empirical Δ_∞ does not by itself certify ε-DP.* On UCI Adult / RandomForest the empirical $\Delta_\infty/g \ll 1$ regime makes the mechanism near-perfect at low ε; on German Credit / GradientBoosting $\Delta_\infty/g \approx 9$ and the mechanism is not deployable. See §8 for cross-dataset numbers and the configuration-dependent reconciliation of the Adult-RF results.

Negative result for full-vector release (empirical). The full-vector Gaussian mechanism has no practical operating point in our evaluated settings: no parameters we tested give both $\varepsilon \leq 10$ and SNR ≥ 0.5 for any model class. The naive global L_2 sensitivity grows as $O(\sqrt{K}/\varepsilon)$ because one background-record replacement perturbs all K coalition evaluations through μ_D.

Secondary baseline: bootstrap-calibrated full-vector mechanism (§5). A bootstrap-calibrated smooth-sensitivity mechanism reduces σ by approximately 24× for MLP and approximately 48–58× for tree models under the corrected centered-response certified Gaussian baseline (Table 7); its privacy interpretation is bootstrap-distributional and conditional on stated dominance/smoothness assumptions; it is not a certified worst-case DP guarantee (Theorem 29). *Although the σ reduction is large, absolute full-vector top-k utility remains limited under the corrected certified baseline* (top-5 accuracy $\sim$ 25–40% in Table 7); this is part of why ranking, not full-vector release, is the lead contribution. We provide the bootstrap mechanism as a baseline for the $\Delta_\infty/g \gg 1$ regime where ranking is not deployable.

Lower-bound sketch (§7). A fingerprinting-style argument suggests an $\Omega(\sqrt{Kd}/\varepsilon)$ scaling for full-vector cardinal release under the (κ, d)-rank condition and single-feature L_1 adjacency, narrowing rather than closing the gap to known DP-SHAP results.

Privacy parameter clarification. The Gaussian-DP parameter $\delta_G = 10^{-5}$ and the bootstrap failure probability $\alpha = 0.01$ used in §5 are reported separately; we do *not* treat their sum as a single DP δ. The ranking mechanism in §4 is pure ε-DP and does not use δ at all.

Keywords: Differential Privacy; Kernel SHAP; Smooth Sensitivity; Bootstrap; Lower Bounds; Explainable AI; Machine Learning Privacy.

*Portions of this work are the subject of International Patent Application PCT/IB2026/053822. The patent-title language should not be read as a scientific certification claim within this manuscript; see Theorem 29 for the actual scope of the bootstrap mechanism's privacy interpretation.

1 Introduction

Setting. A client queries a machine-learning model and receives a kernel SHAP explanation $\varphi(x) \in \mathbb{R}^d$ — per-feature attribution scores computed by the Lundberg–Lee estimator [Lundberg and Lee, 2017]. Publishing $\varphi(x)$ is risky: the SHAP vector is a near-deterministic function of the private input x, and adversaries can reconstruct inputs from published explanations [Luo et al., 2022]. The standard defense — Gaussian differential privacy [Dwork and Roth, 2014] — fails for kernel SHAP at practical privacy budgets.

Why the Gaussian mechanism fails. Kernel SHAP evaluates the model on K masked coalitions and regresses to obtain $\varphi(x) = R\,y(x)$, where $R = (Z^\top W Z)^{-1} Z^\top W$. Replacing one background record perturbs $\Omega(K)$ of the K coalition evaluations simultaneously. The L_2 norm of the resulting perturbation in y-space grows as $\Theta(\sqrt{K})$. To achieve (ε, δ)-DP, the Gaussian mechanism therefore requires $\sigma = \Theta(\sqrt{K}/\varepsilon)$, whereas SHAP values are typically 0.01–0.1 in magnitude. At $K = 400$ and $\varepsilon = 1$, $\sigma \approx 9$ — two orders of magnitude above the signal. The mechanism destroys utility before achieving privacy.

The structural fix. The regression operator R *cancels* the $\sqrt{K}$ growth: for additive models, $R\,\Delta\tilde{y} = w \odot \Delta\mu$, which has L_2 norm $\le \|w\|_2 \|\Delta\mu\|_2$, independent of K (Lemma 5 applied coordinate-by-coordinate). The local L_2 sensitivity of φ is $\Theta(1)$ in K, not $\Theta(\sqrt{K})$, for every model class we study empirically. A mechanism that exploits this structure via smooth sensitivity (Nissim et al. 2007; NRS) combined with a bootstrap upper-confidence calibration of the smooth-sensitivity envelope follows the bootstrap-calibrated conditional privacy interpretation of Theorem 29 (under the dominance/smoothness assumptions) with σ ratios of 24× for MLP and 48–58× for RF/GB at $\varepsilon = 1$ on the recomputed full-scale certified centered-response baseline (Table 7); the 2.4–4.2× RMSE improvement reported in the preliminary tables (Table 17 for MLP and Table 18 in Appendix E for RF/GB) reflects a preliminary $1/n$-based calibration that does not transfer to discontinuous tree models.

What the privacy guarantee is, and is not. The mechanism we propose is *bootstrap-calibrated* rather than worst-case certified. Specifically, the bootstrap upper-confidence estimate S^*_{boot} is a Bernstein-style upper bound on $\mathbb{E}[\mathrm{LS}^{(1)}(D_k)]$ when D_k is drawn from the bootstrap distribution; it does *not*, on its own, certify the supremum $\sup_{D':d(D,D')=k} \mathrm{LS}^{(1)}(D')$ over all adjacent datasets that NRS smooth sensitivity requires. The resulting conditional calibration interpretation (Theorem 29) therefore depends on explicit unproved dominance/smoothness assumptions relating the bootstrap distribution to the worst case, and we report this as a bootstrap-calibrated heuristic baseline rather than a worst-case DP certificate. Throughout, the Gaussian-DP parameter $\delta_G = 10^{-5}$ and the bootstrap failure probability $\alpha = 0.01$ are kept distinct.

Contributions.

1. **Column-cancellation and empirical local sensitivity** (§3). Closed-form bounds for linear and (under a margin assumption) stump models; empirical $\Theta(1)$-in-K confirmation for MLP, RF, GB across $K = 100 \to 800$.

2. **Differentially private ranking release** — lead positive contribution (§4). The exponential mechanism (and its distributionally equivalent Gumbel-max implementation) is pure ε-DP under any *valid* upper bound Δ_∞; a Laplace report-noisy-max with scale $2\Delta_\infty/\varepsilon$ is a standard ε-DP alternative with comparable utility constants. The empirical Δ_∞/g utility law of Cor. 15 governs accuracy. We *certify* Δ_∞ in closed form for linear and logistic models (Lemmas 9–11); for general models we use a conservative bound under rank conditions (Theorem 12) or a direct empirical estimate that serves as a non-certified utility diagnostic only. Adult RF: $\Delta_\infty/g \approx 4{\cdot}10^{-3}$, near-perfect at $\varepsilon = 1$; German GB: $\Delta_\infty/g \approx 9$, not deployable (Table 10).

3. **Bootstrap-calibrated smooth sensitivity mechanism** — secondary baseline (§5). A model-agnostic full-vector mechanism that beats a certified Gaussian baseline by 24× for MLP and 48–58× for RF/GB in σ scale (Table 7); the privacy interpretation is bootstrap-calibrated and conditional on unproved dominance/smoothness assumptions, not worst-case certified (Theorem 29). Useful when ranking is not deployable for a given (dataset, model) pair.

4. **Fingerprinting-style lower-bound sketch** (§7). A workshop/arXiv-grade proof sketch suggesting $\Omega(\sqrt{Kd}/\varepsilon)$ per-coordinate scaling under the (κ, d)-rank condition and single-feature L_1 adjacency; the constant $c \in [1/8, 1]$ in the argument gives a gap range $[1.9\times, 15\times]$ to Mechanism B. We narrow rather than close the gap to known DP-SHAP results.

2 Background

2.1 Kernel SHAP

Let $f : \mathbb{R}^d \to \mathbb{R}$ be the model and $x \in \mathbb{R}^d$ the query input. Fix a background dataset $D = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathbb{R}^d$ with sample mean $\mu_D = n^{-1} \sum_i x^{(i)}$.

Coalition sampling and regression problem used in this paper. We follow the constrained-regression formulation of Lundberg and Lee [2017]. The empty mask $s = \mathbf{0}$ and the full mask $s = \mathbf{1}$ are *not* sampled into Z; instead they are pinned exactly via two hard constraints,

$$f(\mu_D) \;=\; \varphi_0, \qquad f(x) \;=\; \varphi_0 + \mathbf{1}^\top \varphi,$$

which together implement the efficiency axiom $\sum_j \varphi_j = f(x) - f(\mu_D)$. We then sample K *interior* binary masks $s^{(1)}, \ldots, s^{(K)} \in \{0,1\}^d \setminus \{\mathbf{0}, \mathbf{1}\}$ with Shapley-kernel weights $w_k \propto (d-1) \,/\, \big(\binom{d}{|s^{(k)}|} |s^{(k)}|(d - |s^{(k)}|)\big)$, form the coalition matrix $Z \in \{0,1\}^{K \times d}$ (rows $= s^{(k)}$), and evaluate the masked model on each coalition:

$$y_k(D, x) \;=\; f\big(\, x \odot s^{(k)} + \mu_D \odot (1 - s^{(k)})\big).$$

We center the response by subtracting the empty-coalition value $\bar{y} = f(\mu_D)$, set $\tilde{y}_k = y_k - \bar{y}$, and solve the weighted least-squares problem

$$\min_{\varphi \in \mathbb{R}^d} \sum_{k=1}^{K} w_k \big(\tilde{y}_k - s^{(k)\top} \varphi\big)^2 \quad \text{subject to} \quad \mathbf{1}^\top \varphi = f(x) - f(\mu_D).$$

The unconstrained minimiser of the same weighted objective is

$$\tilde{\varphi}(D, x) \;=\; R\, \tilde{y}(D, x), \qquad R = (Z^\top W Z)^{-1} Z^\top W,$$

which is the operator referenced throughout. The constant offset $\bar{y}$ is absorbed by the centering step (equivalently, by the φ_0 intercept), so R acts only on the centered response and constant shifts of y pass through to a constant shift of φ_0, not of φ. In the non-private setting the efficiency constraint is enforced exactly, e.g. by Lagrangian or by orthogonal projection onto the affine hyperplane $\{\varphi : \mathbf{1}^\top \varphi = f(x) - f(\mu_D)\}$. In the private setting we do *not* apply this projection after adding noise: the hyperplane depends on the private quantity $f(\mu_D)$, so the projection is not post-processing in the DP sense (Remark 25). The released private vector therefore satisfies efficiency in expectation only (Theorem 33).

Scope of Lemma 5. Lemma 5 (column-cancellation) is stated for the unconstrained operator R acting on the centered response $\tilde{y}$. Because the private mechanism returns the unprojected vector, L_2 sensitivity bounds on $R\tilde{y}$ apply directly to the released output.

2.2 Adjacency models used in this paper

The paper uses three distinct adjacency relations in different theorems. We list them explicitly here and indicate which result uses which model; results are not transferred between models without proof.

Definition 1 (Single-record replacement, used by §5)**.** Background datasets D, D' are *adjacent* ($D \sim D'$) if they differ in exactly one record: $|D \triangle D'| = 1$.

Definition 2 (Single-feature L_1 adjacency, used by Thm 40)**.** Inputs (or planted backgrounds) x, x' are adjacent if $\sum_j |x_j - x'_j| \le 1$. This is the model used by the fingerprinting construction.

Table 1: Adjacency model used by each theorem. We do not transfer results across rows without explicit proof.

Result	Adjacency model	What changes
Lemma 5, Prop. 6	single-record replacement (background)	one record in D
Thm 29 (mechanism privacy)	single-record replacement (background)	one record in D
Thm 40 (lower bound)	single-feature L_1 on planted background	one feature, $\|\cdot\|_1 \le 1$
Cor. 41	L_∞ on planted background	all features, $\|\cdot\|_\infty \le 1$
Thm 12 (ranking)	single-feature L_1 on input/background	one feature, $\|\cdot\|_1 \le 1$

Definition 3 (L_∞ adjacency, used by Cor. 41)**.** Inputs x, x' are adjacent if $\|x - x'\|_\infty \le 1$. Each coordinate may move by up to 1 simultaneously.

The mechanism's setting is that of a client publishing $\varphi(x)$ when the background is a population dataset; a single individual's presence or absence in the background changes μ_D, hence all K coalition evaluations. This is distinct from *data-level* adjacency (training-set replacement), studied by Patel et al. [2022], which has lower sensitivity.

Threat-model and certification map. Table 2 consolidates which mechanism/result protects which private object, under which adjacency relation, and whether the privacy claim is a real DP certificate or a heuristic/conditional statement.

Table 2: Threat-model and certification map. "Certified?" indicates whether the corresponding mechanism/result has a real $(\varepsilon[, \delta])$-DP proof in the present manuscript or is reported as a non-certified diagnostic / heuristic / conditional statement.

Mechanism / result	Private object	Adjacency	Certified?
Ranking, linear (Lemma 9, 10) + exp. mech. (Thm 14, Thm 16)	query or background	query/input or single-record replacement	yes (ε-DP)
Ranking, logistic (Lemma 11) + exp. mech. (Thm 14)	query or background	query/input or single-record replacement	yes, conservative
Ranking, general nonlinear empirical	query/background	sampled perturbations	no, diagnostic only (§4.2)
Gaussian full-vector (Thm 19 + Lem 20)	background	single-record replacement	yes (ε, δ)-DP, poor utility
Bootstrap full-vector (Algorithm 1, Thm 29)	background	bootstrap distributional	no; heuristic/conditional baseline only (Prop. 22)
Lower bound (Thm 40)	planted input/background	single-feature L_1	proof sketch

2.3 Differential privacy and the Gaussian mechanism

A randomised mechanism $M : \mathcal{D} \to \mathbb{R}^d$ is (ε, δ)-*DP* if for all adjacent $D \sim D'$ and all $S \subseteq \mathbb{R}^d$: $\Pr[M(D) \in S] \le e^\varepsilon \Pr[M(D') \in S] + \delta$. The Gaussian mechanism adds $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = \Delta_2 \sqrt{2 \ln(1.25/\delta)}/\varepsilon$, where $\Delta_2 = \sup_{D \sim D'} \|f(D) - f(D')\|_2$ is the global L_2 sensitivity [Balle and Wang, 2018].

2.4 NRS smooth sensitivity framework

Nissim et al. [2007] define the β-*smooth sensitivity*:

$$S^*_\beta(D) \;=\; \max_{k \ge 0} e^{-\beta k} \sup_{D' :\, d(D, D') = k} \mathrm{LS}^{(1)}(D'),$$

where $\mathrm{LS}^{(1)}(D') = \sup_{D'' \sim D'} \|\varphi(D') - \varphi(D'')\|_2$ is the local sensitivity at D'. Adding $\mathcal{N}(0, S_\beta^{*2} \cdot \sigma_0^2 I)$ (with σ_0 from the Gaussian calibration) achieves (ε, δ)-DP whenever S_β^* is correctly computed.

2.5 Empirical Bernstein inequality

Lemma 4 (Maurer–Pontil [Maurer and Pontil, 2009])**.** *Let $X_1, \ldots, X_B$ be i.i.d. with values in $[0, M]$. With probability $\geq 1 - \alpha$:*

$$\mathbb{E}[X] \;\leq\; \bar{X}_B + \sqrt{\frac{2\hat{V}_B \ln(2/\alpha)}{B}} + \frac{7M \ln(2/\alpha)}{3(B-1)},$$

where $\bar{X}_B$ and $\hat{V}_B$ are the sample mean and sample variance.

3 Column-Cancellation and Local Sensitivity

3.1 Column-cancellation lemma

Lemma 5 (Column cancellation)**.** *Let $z_i \in \mathbb{R}^K$ denote the i-th column of Z, and assume $M := Z^\top W Z$ is invertible. If $\Delta y = c\, z_i$ for some scalar $c \in \mathbb{R}$, then*

$$\|\Delta\hat{\varphi}\|_2 = \|R(c\, z_i)\|_2 = |c|,$$

independent of K.

Proof. $(Z^\top W z_i)_j = \sum_k w_k Z_{ki} Z_{kj} = M_{ij}$, so $Z^\top W z_i = M e_i$. Therefore $R z_i = M^{-1}(M e_i) = e_i$, and $\|R(c\, z_i)\|_2 = |c| \|e_i\|_2 = |c|$. □

Implication for additive models. For additive $f(x) = w^\top x$, replacing one background record changes μ_D by $\Delta\mu = (\Delta x^{(r)})/n$, shifting each coalition evaluation by $y_k \mapsto y_k + w^\top(s^{(k)} \odot \Delta\mu - \Delta\mu) = y_k - w^\top\Delta\mu + w^\top(s^{(k)} \odot \Delta\mu)$. The term $-w^\top\Delta\mu$ is a constant offset and disappears under the centering of §2.1; the remaining perturbation is $\Delta\tilde{y}_k = (w \odot \Delta\mu)^\top s^{(k)} = \sum_j (w_j \Delta\mu_j) Z_{kj}$, i.e.

$$\Delta\tilde{y} \;=\; \sum_{j=1}^{d} (w_j \Delta\mu_j)\, z_j \;=\; Z\,(w \odot \Delta\mu).$$

Lemma 5 applied coordinate-by-coordinate gives $R\,\Delta\tilde{y} = w \odot \Delta\mu$, so $\|R\,\Delta\tilde{y}\|_2 = \|w \odot \Delta\mu\|_2 \leq \|w\|_2 \|\Delta\mu\|_2$. Column cancellation thus collapses an arbitrary additive-model perturbation to the coordinate-wise product $w \odot \Delta\mu$, not to a single column $c\, z_i$ — the latter holds only when $\Delta\mu$ is supported on a single coordinate. Threshold-stump models are discontinuous; the analogous coordinate-wise structure holds only under the no-crossing margin assumption of Proposition 7.

Proposition 6 (Local sensitivity for additive models)**.** *Let $f(x) = w^\top x$, background D of size n, single-record replacement adjacency, and assume each background record is clipped in L_2: $\|x^{(i)}\|_2 \leq B_{\mathrm{clip}}$ for all i. Then*

$$\mathrm{LS}(\varphi) \;\leq\; \frac{2B_{\mathrm{clip}}\, \|w\|_2}{n}.$$

The bound is K-independent. If instead the assumption is the coordinate-wise L_∞ clip $\|x^{(i)}\|_\infty \leq B_{\mathrm{clip}}$, then the corresponding bound acquires a $\sqrt{d}$ factor: $\mathrm{LS}(\varphi) \leq 2B_{\mathrm{clip}}\sqrt{d}\, \|w\|_2/n$. We use the L_2-clipping form throughout the paper; experiments preprocess inputs by L_2 projection onto the ball of radius B_{clip} (§8). Empirical/closed-form ratio $\in [1.02, 1.76]$ across tested (d, K) configurations under L_2 clipping.

Proposition 7 (Local sensitivity for threshold-stump models)**.** *Let $f(x) = \sum_j u_j \mathbf{1}[x_j > 0]$, background D of size n, and single-record replacement adjacency. Threshold models are discontinuous in x: a single replacement can shift μ_D by $\Theta(1/n)$ in some coordinate, and an arbitrarily small such shift can flip $\mathbf{1}[\mu_{D,j} > 0]$ if the coordinate sits exactly on the threshold.*

Under the margin assumption

$$\min_{j\in[d]} |\mu_{D,j}| \;\geq\; \gamma \quad \textit{with} \quad \gamma > 2B_{\text{clip}}/n,$$

no single-record replacement crosses any threshold, every coalition evaluation is unchanged, $\Delta y = 0$*, and consequently*

$$\text{LS}(\varphi) \;=\; 0.$$

Without the margin assumption *the bound is qualitatively different. A single replacement can shift several coordinates of* μ_D *across their respective thresholds simultaneously. Each crossing on coordinate* j *affects* every *coalition with* $s_j = 0$*; after centering, the resulting perturbation in* $\tilde{y}$ *is a sum of column-aligned terms* $\sum_{j\in\mathcal{F}} u_j\, z_j$*, where* $\mathcal{F}$ *is the set of crossing coordinates. Column cancellation collapses each term to the corresponding* e_j*, so*

$$\text{LS}(\varphi) \;\leq\; \|u\|_2, \qquad \|\Delta\varphi\|_\infty \;\leq\; \|u\|_\infty,$$

both independent of n. *The earlier* $\|u\|_2/n$ *bound stated in a draft was not correct: discontinuity destroys the* $1/n$ *scaling.*

3.2 Local sensitivity for non-additive models

For non-additive f (MLP, tree ensembles), Δy has a component orthogonal to z_i, so Lemma 5 cancels the column-aligned part exactly while the orthogonal part passes through R. Empirically, across $K \in \{100, 200, 400, 800\}$ and $d = 20$:

Table 3: Mean per-explicand local sensitivity $\text{LS}(\varphi)$ vs. K. All values $\Theta(1)$ in K; no growth observed.

Model	$K = 100$	$K = 200$	$K = 400$	$K = 800$
MLP (tanh, $h = 32$)	0.082	0.085	0.085	0.086
RF ($T = 100$, $d_{\max} = 4$)	0.048	0.051	0.050	0.051
GB ($T = 100$, $d_{\max} = 3$)	0.072	0.075	0.075	0.076

Certified upper bound for the Bernstein range. The Bernstein correction (Lemma 4) requires a certified upper bound M_{LS} on the range of the per-bootstrap statistic $\text{LS}_{k,b}$. Algorithm 1 operates on the centered response $\tilde{y}$ (§2.1), so by triangle inequality, sub-multiplicativity, and the centered-response identity

$$\tilde{y}_k(D, x) - \tilde{y}_k(D', x) \;=\; \big[\, y_k(D, x) - y_k(D', x) \,\big] \;-\; \big[\, f(\mu_D) - f(\mu_{D'}) \,\big],$$

each coordinate of $\Delta\tilde{y}$ is bounded by $4F_{\max}$, giving

$$M_{\text{LS}}^{\text{naive}} \;=\; \|R\|_2 \cdot 4F_{\max} \cdot \sqrt{K}, \tag{1}$$

which holds for any model with $|f| \leq F_{\max}$ — including discontinuous tree models. **This is the certified centered-response range we use for tree models, consistent with Lemma 20.**

Tighter $1/n$ **form requires Lipschitz** f**.** For continuous models we can do better. If f is L_f-Lipschitz in the masked-input argument, then $|\Delta y_k| \leq L_f \|\Delta\mu\|_2 \leq 2B_{\text{clip}} L_f/n$ under the L_2-clipping assumption (§3), giving

$$M_{\text{LS}}^{\text{Lip}} \;=\; \|R\|_2 \cdot \frac{2B_{\text{clip}} L_f \sqrt{K}}{n}, \tag{2}$$

provided the model is Lipschitz. For our MLP an explicit Lipschitz constant follows from $L_f \leq \|W_2\|_2 \|W_1\|_2$ (since tanh is 1-Lipschitz). For tree ensembles the model is discontinuous and the Lipschitz form does *not* apply: a $1/n$ shift in μ_D can cross a split threshold and produce an $\Theta(F_{\max})$ jump in y_k. Either an explicit no-threshold-crossing margin assumption on μ_D must be stated, or the mechanism must use the conservative naive form $M_{\text{LS}}^{\text{naive}}$ for tree models.

In our calibration tables and reported σ values for MLP we use $M_{\mathrm{LS}}^{\mathrm{Lip}}$ where justified. For RandomForest and GradientBoosting, *the certified tree-model baseline is Table 7* (recomputed with the naive form (1), no $1/n$). The preliminary $1/n$-based calibrations (Tables 18 and 19 in Appendix E) are retained only for transparency about how the earlier draft was calibrated and *should not be interpreted as certified for discontinuous tree models*. References to M_{LS} in the algorithm and in §8 should be read as $M_{\mathrm{LS}}^{\mathrm{Lip}}$ when an explicit Lipschitz constant is available, and as $M_{\mathrm{LS}}^{\mathrm{naive}}$ otherwise. The M_{LS} value is *not* used as the DP noise calibration; it only appears as the range parameter inside the Bernstein concentration term.

4 Differentially Private Ranking Release: Lead Result

This section is the lead positive contribution of the paper. We give a *differentially private ranking mechanism* for kernel SHAP whose utility is governed by the dimensionless ratio Δ_∞/g between per-coordinate ranking sensitivity and the top-1/top-2 magnitude gap. We provide:

1. a definition of Δ_∞ as the input to the mechanism (Definition 8);
2. *certified* per-coordinate Δ_∞ bounds for linear and logistic models (Lemmas 9–11);
3. a conservative bound under rank conditions for general models (Theorem 12);
4. three deployable ranking mechanisms — report-noisy-max top-1, a top-1+magnitude hybrid, and sparse-thresholded release — together with the empirical Δ_∞/g utility law of Cor. 15;
5. a decision framework organised by the Δ_∞/g regime (Table 5).

The bootstrap-calibrated full-vector mechanism of §5 should be read as a secondary baseline — useful when ranking is not deployable for a given (dataset, model) pair — rather than the primary contribution.

Definition 8 (Per-coordinate ranking sensitivity, by assumption)**.** For a query x, an adjacency relation $\sim$, and a coalition design (Z, W), the per-coordinate ranking sensitivity is

$$\Delta_\infty \;:=\; \sup_{x \sim x'} \big\| \, |\varphi(x)| - |\varphi(x')| \, \big\|_\infty .$$

We treat Δ_∞ as a quantity that must be upper-bounded by a mechanism input, either via a structural certificate (Lemmas 9–11 below for linear and logistic models, the conservative Theorem 12 for general models) or via direct empirical estimation (§8, Appendix D).

4.1 Certified Δ_∞ for linear and logistic models

Two adjacency variants. The ranking mechanism's privacy depends on which adjacency relation we protect against. We give certified Δ_∞ bounds for both:

- **Query/input adjacency.** The query x varies; the background D (and hence μ_D) is held fixed. Adjacent inputs differ in one coordinate j by at most ρ: $|x_j - x'_j| \le \rho$. This is the natural adjacency for clients who repeatedly query a fixed model with similar but adjacent inputs.
- **Background-record-replacement adjacency.** The query x is held fixed; the background D varies by replacing one record. This is the adjacency that the rest of the paper uses for the bootstrap full-vector mechanism (§5). Adjacent backgrounds differ in one record under L_2 clipping $\|x^{(i)}\|_2 \le B_{\mathrm{clip}}$.

Lemma 9 (Linear Δ_∞, query/input adjacency)**.** *Let $f(x) = w^\top x$. Under query/input adjacency where x' differs from x in one coordinate j with $|x_j - x'_j| \le \rho$,*

$$\big\| \, |\varphi(x)| - |\varphi(x')| \, \big\|_\infty \;\le\; \|\varphi(x) - \varphi(x')\|_\infty \;\le\; |w_j| \, \rho,$$

hence $\Delta_\infty^{\mathrm{q}} \le \rho \max_j |w_j|$.

Proof. For the centered response, $\Delta\tilde{y}_k = w^\top(s^{(k)} \odot x - s^{(k)} \odot x') = w_j s_j^{(k)}(x_j - x'_j)$, so $\Delta\tilde{y} = w_j(x_j - x'_j)\, z_j$ is column-aligned. Lemma 5 gives $R\,\Delta\tilde{y} = w_j(x_j - x'_j)\, e_j$, with L_∞ norm $|w_j||x_j - x'_j| \le |w_j|\rho$. The first inequality is the reverse triangle inequality coordinate-wise. □

Lemma 10 (Linear Δ_∞, background-record-replacement adjacency)**.** *Let $f(x) = w^\top x$, background D of size n, and let D' replace one record of D with another satisfying the L_2-clipping assumption $\|x^{(i)}\|_2 \le B_{\text{clip}}$. Then for any fixed query x,*

$$\big\|\, |\varphi(D,x)| - |\varphi(D',x)| \,\big\|_\infty \;\le\; \|\varphi(D,x) - \varphi(D',x)\|_\infty \;\le\; \frac{2B_{\text{clip}}}{n} \max_j |w_j|,$$

hence $\Delta_\infty^{\text{bg}} \le (2B_{\text{clip}}/n)\max_j |w_j|$. This bound carries the $1/n$ factor that the query/input version of Lemma 9 does not, because background-record replacement perturbs μ_D by $O(1/n)$ rather than perturbing x directly.

Proof. Replacing one record changes μ_D by $\Delta\mu = (x'^{(r)} - x^{(r)})/n$, so $\|\Delta\mu\|_2 \le 2B_{\text{clip}}/n$ under L_2 clipping; consequently $|\Delta\mu_j| \le 2B_{\text{clip}}/n$ coordinate-wise. For $f(x) = w^\top x$, the coalition perturbation is $\Delta y_k = w^\top((1 - s^{(k)}) \odot \Delta\mu)$. After centering by the empty-coalition baseline, $\Delta\tilde{y}_k = -s_k^\top(w \odot \Delta\mu) = -\sum_j s_j^{(k)} w_j \Delta\mu_j$, i.e. $\Delta\tilde{y} = -Z(w \odot \Delta\mu)$. Since $RZ = I$ when $Z^\top W Z$ is invertible, $R\,\Delta\tilde{y} = -(w \odot \Delta\mu)$, and $\|R\,\Delta\tilde{y}\|_\infty = \max_j |w_j \Delta\mu_j| \le \max_j |w_j| \cdot 2B_{\text{clip}}/n$. □

Lemma 11 (Conservative logistic Δ_∞ bound via sigmoid Lipschitzness)**.** *Let $f(x) = \sigma(w^\top x)$ with $\sigma(\cdot)$ the logistic sigmoid ($\frac{1}{4}$-Lipschitz).* This is a conservative bound, not a clean direct certificate as in Lemma 9. *Under query/input adjacency on coordinate j with $|x_j - x'_j| \le \rho$, $|\Delta y_k| \le \frac{1}{4}|w_j|\rho$ on every coalition for which the relevant component of the masked input changes, so $\|\Delta y\|_2 \le \frac{1}{4}|w_j|\rho\sqrt{K}$. Applying the conservative Theorem 12 with $L_0 = \frac{1}{4}|w_j|\rho\sqrt{K}$:*

$$\Delta_\infty^{\text{q}} \;\le\; \frac{|w_j|\rho}{4} \cdot \frac{\sqrt{w_{\max}/w_{\min}}}{\kappa\sqrt{K/d}} \cdot \sqrt{K} \;=\; \frac{|w_j|\rho}{4} \cdot \frac{\sqrt{d\,(w_{\max}/w_{\min})}}{\kappa}.$$

For background-record-replacement adjacency we derive the bound independently, without inheriting the linear $RZ = I$ cancellation of Lemma 10: replacing one background record changes μ_D by $\Delta\mu = (x'^{(r)} - x^{(r)})/n$ with $\|\Delta\mu\|_2 \le 2B_{\text{clip}}/n$. For each coalition k, the masked-input argument shifts only on the unmasked coordinates, so the pre-sigmoid input $w^\top(\cdot)$ shifts by at most $\|w\|_2 \cdot 2B_{\text{clip}}/n$. Sigmoid is $1/4$-Lipschitz, so $|\Delta y_k| \le (1/4)\|w\|_2 \cdot 2B_{\text{clip}}/n$ for every k, giving

$$\|\Delta y\|_2 \;\le\; \frac{\|w\|_2\, B_{\text{clip}}\sqrt{K}}{2n}.$$

Plugging this into the operator-norm bound $\|R\,\Delta\tilde{y}\|_\infty \le \|R\|_{2\to\infty} \cdot \|\Delta y\|_2$ of Theorem 12 (which gives $\|R\|_{2\to\infty} \le \sqrt{w_{\max}/w_{\min}}/(\kappa\sqrt{K/d})$ under the rank condition):

$$\Delta_\infty^{\text{bg}} \;\le\; \frac{\|w\|_2\, B_{\text{clip}}\sqrt{d\,(w_{\max}/w_{\min})}}{2n\,\kappa}.$$

This bound uses sigmoid Lipschitzness on the scalar pre-activation together with the rank-condition operator-norm bound; it does not *invoke the exact column cancellation of Lemma 10, which holds only for $f(x) = w^\top x$. Both bounds are loose: the empirical Adult logistic regime gives $\Delta_\infty/g \ll 1$ at modest ε in our experiments, far better than these conservative estimates suggest.*

Theorem 12 (Conservative bound on Δ_∞ for general models)**.** *Assume single-feature L_1 adjacency $\|x - x'\|_1 \le 1$, the (κ, d)-rank condition on Z with $\kappa \ge 1/2$, and that the* coalition-evaluation map *$x \mapsto y(D, x, Z) \in \mathbb{R}^K$ is L_0-Lipschitz in the perturbed coordinate, so $\|\Delta y\|_2 \le L_0$ — note that for kernel SHAP, L_0 is the Lipschitz constant of the entire K-vector evaluation, not of the scalar model output, and may scale with $\sqrt{K}$ in the worst case. Let $w_{\max}$ and $w_{\min}$ be the largest and smallest Shapley-kernel weight over the sampled interior coalitions. Then*

$$\Delta_\infty \;\le\; \|R\|_{2\to\infty} \cdot L_0 \;\le\; \frac{L_0\sqrt{w_{\max}/w_{\min}}}{\kappa\sqrt{K/d}}.$$

This is the bound we prove in Appendix C. The proof relies on a singular-value bound on $\sigma_d(Z)$, which gives a $2 \to 2$ inverse-norm control; converting this to the desired $2 \to \infty$ control inside $\|R\|_{2\to\infty}$ is the source of the $\sqrt{w_{\max}/w_{\min}}$ factor.

Remark 13 (The clean $L_0/\sqrt{d}$ form is a working assumption)**.** The cleaner statement $\Delta_\infty \le L_0/\sqrt{d}$ used informally in earlier drafts is *not* implied by Theorem 12 under the listed hypotheses alone. It requires either (i) a weighted-coherence assumption stronger than what binary masks supply (e.g. that $\|W^{1/2}Z\|_{2\to\infty}\sqrt{K} \le c$ for an absolute constant c, equivalent to each row of $W^{1/2}Z$ being $O(1)$-sparse rather than $O(d)$-sparse), or (ii) an explicit ratio bound $w_{\max}/w_{\min} = O(1)$, which the Shapley kernel does not in general provide across all coalition sizes, or (iii) a direct empirical estimate of Δ_∞ from the actual (Z, W, f, D) in use. We adopt option (iii) for non-linear/non-logistic models: the ranking utility numbers in §8 for MLP, RandomForest, and GradientBoosting are computed by plugging an empirically estimated Δ_∞ into the privacy and accuracy formulas, not by invoking a closed-form $L_0/\sqrt{d}$ bound. For linear and logistic models we use the certified bounds of Lemmas 9–11 instead.

4.2 Top-1 release via the exponential mechanism (main theorem)

This subsection contains the lead certified result of the paper. Given any valid upper bound Δ_∞ on the per-coordinate ranking sensitivity (Definition 8 — Lemma 9 or Lemma 10 for linear models, Lemma 11 for logistic, conservative analytic bound from Theorem 12 for general models, or an externally certified bound), we release the top-1 feature via the *exponential mechanism*.

Theorem 14 (Pure-DP top-1 SHAP ranking release)**.** *Fix a query x, background dataset D, coalition design (Z, W), and attribution vector $\varphi(D, x) \in \mathbb{R}^d$. Let*

$$\Delta_\infty \;\ge\; \sup_{D\sim D'} \big\| \,|\varphi(D,x)| - |\varphi(D',x)|\, \big\|_\infty$$

be a valid upper bound on the per-coordinate ranking sensitivity under the chosen adjacency relation. Define the mechanism

$$\Pr[M(D) = i] \;=\; \frac{\exp\!\big(\varepsilon\,|\varphi_i(D,x)|/(2\Delta_\infty)\big)}{\sum_{j=1}^{d}\exp\!\big(\varepsilon\,|\varphi_j(D,x)|/(2\Delta_\infty)\big)}.$$

Then M is ε-differentially private.

Proof. Define the quality score $q(D, i) = |\varphi_i(D, x)|$. For adjacent $D \sim D'$ and any $i \in [d]$, the reverse triangle inequality and the definition of Δ_∞ give

$$|q(D,i) - q(D',i)| \;=\; \big|\,|\varphi_i(D,x)| - |\varphi_i(D',x)|\,\big| \;\le\; \big\|\,|\varphi(D,x)| - |\varphi(D',x)|\,\big\|_\infty \;\le\; \Delta_\infty.$$

Thus the score q has global sensitivity at most Δ_∞. The exponential mechanism with score sensitivity Δ_∞ samples item i with probability proportional to $\exp(\varepsilon\, q(D, i)/(2\Delta_\infty))$, and is ε-DP by the standard exponential-mechanism theorem [Dwork and Roth, 2014, Theorem 3.10]. □

A *Gumbel-max* implementation is distributionally equivalent to the exponential mechanism — i.e. samples from the same distribution on i^*. A *Laplace report-noisy-max* implementation with scale $2\Delta_\infty/\varepsilon$ is a standard ε-DP alternative with comparable utility constants but is *not* distributionally identical to the exponential mechanism: the noisy-max distribution under Laplace noise differs from the exponential mechanism's softmax, and the equivalence is at the level of utility bounds up to constants, not exact sample distribution. We use Gumbel-max when the formal exponential-mechanism distribution is required and Laplace RNM when only the standard ε-DP guarantee and comparable constants are needed. **Empirically deployable when $\Delta_\infty/g \ll 1$ at the chosen ε and Δ_∞ is a valid (analytic or externally certified) upper bound (e.g. Adult RF in our experiments).**

Empirical Δ_∞ is a utility diagnostic, not a DP certificate. The values of Δ_∞ reported for general models in §8 (Tables 9 and 10) are computed as the worst-case $\big\||\varphi(x)| - |\varphi(x')|\big\|_\infty$ over a finite set of random adjacent inputs. They are *non-certified* estimates of Δ_∞ and serve only as predictors of mechanism utility via Cor. 15. The pure ε-DP guarantee of the exponential mechanism / Laplace RNM holds only when the Δ_∞ used for noise calibration is itself a valid *upper bound* on the true per-coordinate ranking sensitivity — i.e. derived analytically (Lemmas 9, 10, 11, Theorem 12) or supplied as a certified bound by an external analysis. An empirical estimate computed from a finite perturbation set may underestimate the true sensitivity and therefore must *not* be used as the noise-calibration parameter in a deployed DP release. In §8 we use empirical Δ_∞ only as a non-certified utility diagnostic to predict accuracy, not to claim DP for the underlying mechanism on those configurations.

Corollary 15 (Top-1 correctness)**.** *Let $i^\star = \arg\max_i |\varphi_i(D,x)|$ and let $g = |\varphi_{i^\star}(D,x)| - \max_{j\neq i^\star} |\varphi_j(D,x)|$ denote the top-1/top-2 magnitude gap. Then the exponential mechanism of Theorem 14 satisfies*

$$\Pr[M(D) = i^\star] \;\geq\; 1 - d\exp\!\left(-\frac{\varepsilon g}{2\Delta_\infty}\right).$$

Proof. For every incorrect feature $j \neq i^\star$, $q(D,i^\star) - q(D,j) \geq g$ by definition of g. The standard exponential-mechanism utility bound [Dwork and Roth, 2014, Theorem 3.11] gives

$$\Pr\big[q(D,M(D)) \leq q(D,i^\star) - g\big] \;\leq\; d\exp\!\left(-\frac{\varepsilon g}{2\Delta_\infty}\right).$$

The event on the left contains the event that the selected feature is not a top-1 maximiser (up to ties), so $\Pr[M(D) = i^\star] \geq 1 - d\exp(-\varepsilon g/(2\Delta_\infty))$. □

Reading the law. Utility is governed entirely by the dimensionless ratio Δ_∞/g: when $\Delta_\infty/g \ll 1$ the mechanism is near-perfect at low ε; when $\Delta_\infty/g \gg 1$ it is not deployable without additional structure (Table 10). We do *not* assert a fixed ε at which 90% accuracy is reached; that threshold is configuration-dependent.

RNM (Laplace) implementation note. A common implementation is report-noisy-max: $i^* = \arg\max_i\big(|\varphi_i| + \mathrm{Lap}(0, 2\Delta_\infty/\varepsilon)\big)$. With the Laplace scale $2\Delta_\infty/\varepsilon$ this is ε-DP by the standard noisy-max analysis and yields utility bounds with the same asymptotic dependence on Δ_∞/g as the exponential mechanism above (constants differ; the two mechanisms are not distributionally identical). Earlier drafts of this paper used the scale $\Delta_\infty/\varepsilon$, which corresponds to a different (more aggressive) calibration; we have switched to the $2\Delta_\infty/\varepsilon$ form for consistency with the formal theorem.

4.3 Hybrid: top-1 + noisy magnitude

Split the privacy budget: (i) release i^* via RNM at budget ε_1; (ii) release $|\varphi_{i^*}|$ via the Gaussian mechanism at per-coordinate sensitivity (scalar, no $\sqrt{K}$) at budget ε_2. Total cost: $(\varepsilon_1 + \varepsilon_2, \delta)$-DP by basic composition. The top-1 correctness probability follows Cor. 15 at budget ε_1; the magnitude SNR depends on $|\varphi_{i^*}|$ relative to the per-coordinate Gaussian noise at ε_2. **Empirically deployable in configurations where $\Delta_\infty/g \lesssim 1$ at the chosen ε_1.**

4.4 Top-k ranking via sequential exponential mechanism

Top-1 release extends to top-k release by sequentially applying the exponential mechanism without replacement.

Theorem 16 (Pure-DP top-k SHAP ranking release)**.** *Fix $k \in \{1,\ldots,d\}$, a query x, background dataset D, and a valid upper bound Δ_∞ on the per-coordinate ranking sensitivity (Definition 8). Iterate for $\ell = 1,\ldots,k$: at step ℓ, run the exponential mechanism of Theorem 14 restricted to the unselected coordinates, with budget ε/k and sensitivity Δ_∞, and append the selected feature to the output set S. Then the resulting mechanism M_k outputting $S \subseteq [d]$, $|S| = k$, is ε-DP.*

Proof. At step ℓ, condition on the previously selected set $S_{\ell-1} \subseteq [d]$ (which is part of the mechanism's public output) and run the exponential mechanism over the finite candidate set $[d] \setminus S_{\ell-1}$ with score $q(D, i) = |\varphi_i(D, x)|$. For any fixed $S_{\ell-1}$, q has score sensitivity at most Δ_∞ on the restricted candidate set by the same argument as in Theorem 14. Hence each step is (ε/k)-DP by the standard exponential-mechanism theorem [Dwork and Roth, 2014, Theorem 3.10]. Adaptive composition [Dwork and Roth, 2014, Theorem 3.16] over the k calls — adaptive because the candidate set at step ℓ depends on the public output of steps $1, \ldots, \ell - 1$ — gives total privacy cost $k \cdot (\varepsilon/k) = \varepsilon$. □

Remark 17 (Utility under top-k composition)**.** Each step ℓ inherits the Cor. 15 bound at the reduced budget ε/k, so the per-step error probability scales as $d \exp(-\varepsilon g_\ell/(2k\Delta_\infty))$ with g_ℓ the gap between the largest unselected magnitude and the next-largest unselected magnitude. Top-k utility is governed by the worst per-step ratio Δ_∞/g_ℓ and the budget split ε/k. An alternative allocation — assigning unequal budgets across steps based on observed gaps — is straightforward but data-adaptive and would require spending an additional budget on the gap estimates if the allocation is to remain DP; we do not pursue this here.

Applicability beyond Kernel SHAP. Theorem 14 (and Theorem 16) applies to any explanation vector $a(D, x) \in \mathbb{R}^d$ for which a valid per-coordinate ranking sensitivity bound is available. Kernel SHAP enters this paper through the derivation and estimation of Δ_∞ (Lemmas 9–11, Theorem 12, the empirical diagnostic of §4.2) and through the local-sensitivity story of §3. The exponential-mechanism layer is independent of how a is computed; analogous certified bounds for TreeSHAP, gradient explanations, integrated gradients, and other per-feature attribution vectors would yield the same ε-DP top-k release mechanism with utility law governed by the corresponding empirical Δ_∞/g ratio.

4.5 Sparse-thresholded release

Release $\{(i, \tilde{\varphi}_i) : |\varphi_i| > \tau\}$ with public threshold τ. *Under a support-stability assumption* — namely that adjacent inputs change the thresholded support $\{i : |\varphi_i| > \tau\}$ by at most one coordinate — the count sensitivity is 1. This support-stability assumption holds in linear/additive settings where the column-cancellation structure limits each adjacent perturbation to a single attribution coordinate (Lemma 5); it does *not* hold in general for non-linear models, where a single input-coordinate perturbation can change multiple attribution coordinates through interactions and shift several across the threshold simultaneously. For non-linear models the count sensitivity should be bounded by an explicit analysis (worst-case number of threshold crossings induced by an adjacent input) or empirically calibrated and reported as a non-certified diagnostic. Propose-test-release [Dwork et al., 2009] releases noisy magnitudes when the support is stable. **Empirically deployable at** $\varepsilon \leq 5$ for stable-support inputs in our evaluated settings.

4.6 Failure-case taxonomy

When the ranking mechanism fails to deliver a useful top-k release, the cause typically falls into one of five categories. We list them explicitly so that diagnosis can be done from the empirical ratio Δ_∞/g and the chosen adjacency model:

1. **Small-gap failure.** g is too small (close to a tie), so the top-1 feature is not semantically stable even non-privately. Non-private top-1 is itself unreliable here; private release should be conditioned on g exceeding a stability threshold (see Remark 18 below) or on top-k release (Theorem 16).

2. **High-sensitivity failure.** Δ_∞ is too large relative to the available budget. For the linear/logistic certified bounds the only fix is reducing the perturbation magnitude ρ or using a tighter analytic certificate.

3. **Uncertified-sensitivity failure.** For nonlinear models, the empirical Δ_∞ used for utility prediction is not a valid DP certificate (§4.2); the mechanism is simply not deployable until an analytic upper bound is supplied.

4. **Full-vector noise failure.** The certified Gaussian baseline (Theorem 19) has σ much larger than typical SHAP magnitudes (§8); per-feature release is destroyed.

5. **Adjacency-mismatch failure.** A bound proven under one adjacency relation (e.g. query/input) is misapplied to another (e.g. background-record replacement). Table 1 is the single source of truth; results are not transferred between rows without a separate proof.

Remark 18 (Tie / near-tie diagnostic)**.** When g is close to zero, the top-1 feature is not semantically stable even before privacy noise. We therefore recommend reporting results both overall and conditioned on g exceeding a small stability threshold. Table 11 reports Δ_∞/g, which serves as a proxy for near-tie or high-sensitivity regimes but does not by itself distinguish the two causes: a large ratio may come from small g (near-tie), large Δ_∞ (high sensitivity), or both. A directly computed near-tie fraction (e.g. fraction of queries with $g < 10^{-2}$) is listed as future work in §10; in the present benchmark configurations whose median ratio exceeds ~ 5 should be re-run with top-k release ($k \geq 3$, Theorem 16) or sparse-thresholded release.

4.7 Composition across queries

For pure-DP accounting, the privacy losses compose additively, so T releases at per-query budget ε cost $T\varepsilon$ in total [Dwork and Roth, 2014, Theorem 3.16]. If one is willing to move to approximate DP, advanced composition or privacy-accountant methods (e.g. moments / Rényi accountants) can trade an additional δ for a smaller effective ε over many queries; we do not use that relaxation here, and the budgets reported in this paper are pure-ε accounting.

Table 4: Per-query budget allocation under basic composition for the ranking mechanism. Total budget is fixed at $\varepsilon_{\text{total}} = 1$.

Queries T	Per-query ε	Total ε
1	1.00	1.0
5	0.20	1.0
10	0.10	1.0
50	0.02	1.0

4.8 Decision framework

Table 5: Release mode decision framework, organised by the empirical ratio Δ_∞/g between per-coordinate ranking sensitivity and top-1/top-2 gap. ε is the single-release budget; composition multiplies it across queries. Concrete values of Δ_∞/g for RandomForest and GradientBoosting on UCI Adult and German Credit are reported in Table 10.

Regime	Expected RNM utility	Recommendation
$\Delta_\infty/g \ll 1$	near-perfect at $\varepsilon \approx 1$	RNM top-1 / hybrid (e.g. Adult RF)
$\Delta_\infty/g \approx 1$	mid- to high-ε needed	RNM at $\varepsilon \geq 3$, or sparse-thresholded if support stable
$\Delta_\infty/g \gg 1$	not deployable as RNM	Empirical bootstrap full-vector baseline (§5; not certified), or derive a refined certified coordinate bound
Bootstrap full vector	$\sim 24\times$ (MLP), ~ 48–$58\times$ (RF/GB) over centered-response cert. Gaussian (Tab. 7)	bootstrap-calibrated, not certified worst-case DP
Full-vector Gaussian	needs $\varepsilon \gg 10$ for SNR ≥ 0.5	Not recommended

5 Bootstrap-Calibrated Smooth Sensitivity Mechanism (secondary baseline)

5.1 Certified full-vector Gaussian baseline (Mechanism A)

We first state the certified full-vector baseline that Algorithm 1 compares against. Unlike the bootstrap mechanism, this baseline admits a real (ε, δ)-DP proof.

Theorem 19 (Certified full-vector Gaussian baseline)**.** *Let*

$$\Delta_2 \;\geq\; \sup_{D \sim D'} \|\varphi(D, x) - \varphi(D', x)\|_2$$

be a valid global L_2 sensitivity bound for the full SHAP vector under the chosen adjacency relation. The mechanism

$$M(D) \;=\; \varphi(D, x) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d),$$

with $\sigma = \Delta_2 \sqrt{2 \ln(1.25/\delta)}/\varepsilon$ is (ε, δ)-DP for $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$.

Proof. This is the standard Gaussian mechanism [Dwork and Roth, 2014, Theorem A.1]. For all adjacent $D \sim D'$, $\|\varphi(D, x) - \varphi(D', x)\|_2 \leq \Delta_2$. Adding spherical Gaussian noise with the stated scale therefore satisfies (ε, δ)-DP by the classical Gaussian-mechanism calibration. □

Lemma 20 (Naive model-agnostic full-vector sensitivity)**.** *Assume $|f(u)| \leq F_{\max}$ for all masked inputs u encountered by the coalition design (Z, W) and for the background means μ_D encountered under the adjacency relation. Let $\tilde{y}_k(D, x) = y_k(D, x) - f(\mu_D)$ denote the centered response used by §2.1. Then*

$$\Delta_2 \;\leq\; 4 F_{\max} \sqrt{K}\, \|R\|_2.$$

Proof. For adjacent $D \sim D'$, the centered-response coordinate difference is

$$\tilde{y}_k(D, x) - \tilde{y}_k(D', x) \;=\; \big[\, y_k(D, x) - y_k(D', x) \,\big] \;-\; \big[\, f(\mu_D) - f(\mu_{D'}) \,\big].$$

Since $|f| \leq F_{\max}$, both brackets have absolute value at most $2F_{\max}$, so $|\tilde{y}_k(D, x) - \tilde{y}_k(D', x)| \leq 4F_{\max}$ for every k, and therefore $\|\tilde{y}(D, x) - \tilde{y}(D', x)\|_2 \leq 4F_{\max}\sqrt{K}$. By sub-multiplicativity,

$$\|\varphi(D, x) - \varphi(D', x)\|_2 \;=\; \|R\big(\tilde{y}(D, x) - \tilde{y}(D', x)\big)\|_2 \;\leq\; 4F_{\max}\sqrt{K}\, \|R\|_2.$$

The factor of 4 (rather than 2) is needed because the centered response subtracts the data-dependent baseline $f(\mu_D)$, which is itself perturbed under adjacency; we do not assume R annihilates the constant baseline-shift term across adjacent datasets. □

Theorem 19 together with Lemma 20 (and, where it applies, the tighter Lipschitz form $\Delta_2 \leq 2B_{\text{clip}} L_f \sqrt{K}\, \|R\|_2/n$ for Lipschitz models, derived analogously from Proposition 6) yields the certified Mechanism A baseline used throughout §8. The $\sqrt{K}$ dependence in Δ_2 is the source of the impractical σ scale at deployable ε, motivating the bootstrap mechanism below as a heuristic baseline that is not itself certified.

Remark 21 (Centered-response factor in Table 7)**.** The RF/GB rows of Table 7 use the centered-response $4F_{\max}$ bound from Lemma 20 rather than the $2F_{\max}$ uncentered bound used in early drafts. Relative to those drafts, the certified Mechanism A σ_A values for RF and GB are doubled (e.g. $\sigma_A^{\varepsilon=1} = 293.92$ for RF and 1024.04 for GB versus the prior 146.96 and 512.02). The bootstrap σ_B values are unchanged because Algorithm 1 already operates in post-regression centered space; consequently the certified-baseline ratios σ_A/σ_B rise from $\sim$ 24–29× to $\sim$ 48–58× for tree models. The MLP entries (which use the certified Lipschitz form rather than the naive $F_{\max}$ form) are unaffected, so the MLP ratio remains 24.12×.

5.2 Bootstrap-calibrated Algorithm 1: heuristic, not certified

Proposition 22 (Status of Algorithm 1)**.** *Algorithm 1 is presented as a* bootstrap-calibrated heuristic baseline*, not as a certified (ε, δ)-DP mechanism. The bootstrap distributional upper envelope $S^*_{\text{boot}}(D)$ produced by the algorithm estimates the bootstrap-distribution mean of the local sensitivity (Lemma 4, Remark 27); it does not, on its own, certify the worst-case adversarial supremum $\sup_{D':d(D,D')=k} \text{LS}^{(1)}(D')$ that NRS smooth sensitivity requires, and the realised scale need not be β-smooth across adjacent datasets. Theorem 29 therefore depends on two unproved auxiliary assumptions* (A1)–(A2)*; under those assumptions it is a conditional calibration statement, not a DP theorem. Converting Algorithm 1 into a real (ε, δ)-DP mechanism would require replacing S^*_{boot} by a deterministic certified envelope $S^\dagger_\beta(D)$ (Lemma 23) for which analytic worst-case bounds are computed for the model family in use (open problem (v) in §10).*

Lemma 23 (Smoothness of the certified envelope $S^\dagger_\beta$)**.** *Define the deterministic envelope*

$$S^\dagger_\beta(D) \;=\; \max_{k\geq 0} e^{-\beta k} \sup_{D':d(D,D')\leq k} \text{LS}^{(1)}(D').$$

Then for adjacent $D \sim D'$, $S^\dagger_\beta(D') \leq e^\beta\, S^\dagger_\beta(D)$.

Proof. For any $k \geq 0$ and any E with $d(E, D') \leq k$, the triangle inequality for dataset distance gives $d(E, D) \leq k + 1$. Hence

$$\sup_{E:d(E,D')\leq k} \text{LS}^{(1)}(E) \;\leq\; \sup_{E:d(E,D)\leq k+1} \text{LS}^{(1)}(E),$$

so

$$e^{-\beta k} \sup_{E:d(E,D')\leq k} \text{LS}^{(1)}(E) \;\leq\; e^\beta\, e^{-\beta(k+1)} \sup_{E:d(E,D)\leq k+1} \text{LS}^{(1)}(E) \;\leq\; e^\beta\, S^\dagger_\beta(D).$$

Taking the maximum over k proves $S^\dagger_\beta(D') \leq e^\beta\, S^\dagger_\beta(D)$. □

Lemma 23 is the missing piece for a real DP proof of a smooth-sensitivity-based mechanism: a certified (ε, δ_G)-DP Gaussian smooth-sensitivity release at scale $S^\dagger_\beta(D)\sqrt{2\ln(1.25/\delta_G)}/\varepsilon$ would follow from Lemma 31 with the verified admissibility constants. The hard part is computing or bounding $S^\dagger_\beta(D)$ non-trivially for nonlinear Kernel SHAP, which we leave open (§10, open problem (v)). Until then, Algorithm 1 should be read as a utility baseline only.

5.3 Setup

We operate in the *post-regression* space, applied to the *centered* response $\tilde{y}(D, x, Z) = y(D, x, Z) - f(\mu_D)\mathbf{1}$ as defined in §2.1. The adjacency is single-record replacement in the background D of size n. Throughout this section, every occurrence of $R\,y(\cdot)$ in the algorithm and the local-sensitivity estimator should be read as $R\,\tilde{y}(\cdot)$: R is applied to the centered response, with constant offsets absorbed by the φ_0 intercept and the efficiency constraint held only in expectation, since the private mechanism does not project. The mechanism: (1) computes the true post-regression SHAP output $\varphi_{\text{ref}} = R\,\tilde{y}(D, x, Z)$; (2) estimates S^*_β via B bootstrap samples at each hop count $k = 0, \ldots, K_{\max}$; (3) applies an empirical Bernstein upper-confidence correction; (4) adds calibrated Gaussian noise and releases the unprojected result. Efficiency holds in expectation only (Theorem 33, Remark 25).

Remark 24 (Bootstrap sampling convention)**.** Algorithm 1 uses the standard *with-replacement empirical bootstrap*: each replaced record in D_k is drawn i.i.d. with replacement from the empirical distribution of D. We do *not* use without-replacement subsampling, sampling from a public holdout, or sampling from a synthetic reference distribution; those are valid alternatives but would change the meaning of the Bernstein target $\mathbb{E}_{D_k\sim\mathcal{B}_k}[\text{LS}^{(1)}(D_k)]$ in Lemma 4 and shift the dominance assumption *(A1)* of Theorem 29. Earlier draft phrasing referring to "without-replacement bootstrap from the same population" is corrected to the with-replacement form here.

Algorithm 1 Bootstrap-calibrated smooth sensitivity for Kernel SHAP

Require: model f, background D (size n), query x, coalition matrix Z, regression operator R, certified $F_{\max}$, hop count $K_{\max}$, bootstrap size B_{boot}, confidence level α, privacy budget (ε, δ). All occurrences of $\tilde{y}(\cdot)$ below denote the centered response $y(\cdot) - f(\mu_\cdot)\mathbf{1}$ as in §2.1.

$\varphi_{\text{ref}} \leftarrow R\,\tilde{y}(D, x, Z)$
for $k = 0, 1, \ldots, K_{\max}$ **do**
  for $b = 1, \ldots, B_{\text{boot}}$ **do**
    Sample D_k by replacing k records of D, drawn i.i.d. *with replacement* from the empirical distribution of D (standard with-replacement empirical bootstrap; see Remark 24)
    $\text{LS}_{k,b} \leftarrow \max_r \|R\,\tilde{y}(D_k, x, Z) - R\,\tilde{y}(D_k^{-r}, x, Z)\|_2$ ▷ single-swap local sensitivity at D_k
  end for
  $\hat{\mu}_k \leftarrow \frac{1}{B_{\text{boot}}} \sum_b \text{LS}_{k,b}$; $\hat{V}_k \leftarrow \frac{1}{B_{\text{boot}}-1} \sum_b (\text{LS}_{k,b} - \hat{\mu}_k)^2$
  $\alpha' \leftarrow \alpha/(K_{\max}+1)$ ▷ per-k Bernstein confidence after union bound; see Lemma 4 and Step 2 of Appendix A.2
  $\text{corr}_k \leftarrow \sqrt{\frac{2\hat{V}_k \ln(2/\alpha')}{B_{\text{boot}}}} + \frac{7M_{\text{LS}} \ln(2/\alpha')}{3(B_{\text{boot}}-1)}$ ▷ M_{LS} from (1) or (2)
  $\text{UB}_k \leftarrow \max_b \text{LS}_{k,b} + \text{corr}_k$ ▷ bootstrap distributional upper envelope at hop k, not an adversarial supremum (Remark 27)
end for
$\beta \leftarrow \varepsilon \,/\, (2\ln(1/\delta_G))$
$S^*_{\text{boot}} \leftarrow \max_{k=0}^{K_{\max}} e^{-\beta k} \cdot \text{UB}_k$
$\sigma \leftarrow S^*_{\text{boot}} \cdot \sqrt{2\ln(1.25/\delta_G)} \,/\, \varepsilon$
$\tilde{\varphi} \leftarrow \varphi_{\text{ref}} + \mathcal{N}(0, \sigma^2 I_d)$
return $\tilde{\varphi}$ ▷ do *not* project; see Remark 25

Remark 25 (Why we do *not* project onto the efficiency hyperplane)**.** An earlier version of this mechanism applied the projection $\Pi_{\Sigma=0}(\tilde{\varphi})$ onto the affine hyperplane $\{\varphi : \mathbf{1}^\top \varphi = f(x) - f(\mu_D)\}$ as a final step, arguing that a projection is post-processing of a DP output. *This is incorrect under input-level adjacency.* The hyperplane itself depends on the private background D through $f(\mu_D)$, so the projection map is not a function of the noisy output alone — it accesses the private dataset a second time and would publish $f(\mu_D)$ exactly (since $f(x)$ is known to the client). The post-processing theorem does not apply to such a data-dependent transformation. We therefore release the unprojected noisy vector $\tilde{\varphi}$. The efficiency axiom holds only *in expectation* (Theorem 33); enforcing it exactly would require either (i) a public background dataset, or (ii) spending a separate DP budget on $f(\mu_D)$ via a scalar Gaussian and projecting onto the noisy-sum hyperplane, accounted under composition. We list the latter as an open problem in §10.

Remark 26 ($k^* = 0$ simplifies deployment)**.** In practice, $k^* := \arg\max_k e^{-\beta k} \cdot \text{UB}_k = 0$ for all model families and all deployable $\varepsilon \le 5$: the smooth-sensitivity maximum is always attained at the unperturbed background. Intuitively, for tree ensembles the local sensitivity does not grow when additional records are swapped (ensemble averaging is stabilising), and for MLP the Bernstein upper bounds decay faster than $e^{\beta k}$ grows. When $k^* = 0$, Algorithm 1 reduces to a single-hop bootstrap ($K_{\max} = 0$ suffices): compute $B_{\text{boot}} = 400$ single-swap sensitivities at D itself, apply the Bernstein correction, and calibrate σ. No multi-hop sampling is required.

5.4 Privacy guarantee: what the mechanism does and does not certify

What the bootstrap actually bounds. NRS smooth sensitivity requires the worst-case quantity

$$S^*_\beta(D) \;=\; \max_{k \ge 0} e^{-\beta k} \sup_{D' : d(D,D') = k} \text{LS}^{(1)}(D').$$

Algorithm 1 replaces the inner sup over adjacent datasets at distance k by a *bootstrap distributional upper envelope*: the sample maximum $\max_b \text{LS}_{k,b}$ plus an empirical Bernstein upper-confidence correction that bounds $\mathbb{E}_{D_k \sim \mathcal{B}_k}[\text{LS}^{(1)}(D_k)]$, where $\mathcal{B}_k$ is the bootstrap distribution at hop k. The sample maximum is conservative relative to the bootstrap mean, but is *not* a worst-case supremum: it characterises the

distributional envelope under $\mathcal{B}_k$, not the adversarial envelope required by NRS. We therefore call the resulting quantity a "bootstrap distributional upper envelope" rather than a "smooth-sensitivity upper bound," and the conditional calibration interpretation of Theorem 29 requires the additional dominance assumption *(A1)* to bridge the two notions.

Remark 27 (Meaning of $\max_b \mathrm{LS}_{k,b} + \mathrm{corr}_k$)**.** The quantity $\mathrm{UB}_k = \max_b \mathrm{LS}_{k,b} + \mathrm{corr}_k$ combines (i) the sample maximum over B_{boot} bootstrap draws and (ii) a Bernstein upper-confidence correction targeting the bootstrap-distribution mean. Item (i) is conservative relative to the sample mean but does *not* on its own estimate $\sup_{D':d(D,D')=k} \mathrm{LS}^{(1)}(D')$. We refer to UB_k as a bootstrap distributional upper envelope at hop k; it becomes a worst-case smooth-sensitivity bound only under the dominance assumption *(A1)* of Theorem 29.

Remark 28 (Bootstrap dominance assumption)**.** The bootstrap-calibrated guarantee in Theorem 29 requires the assumption that, with probability $\geq 1 - \alpha_{\mathrm{dom}}$ over the algorithm's randomness, S^*_{boot} produced by Algorithm 1 dominates the worst-case smooth sensitivity $S^*_\beta(D)$. This assumption is non-trivial: the bootstrap distribution may underestimate the supremum if the worst-case adjacent dataset is rare under the empirical distribution of D. We treat this as a working assumption rather than a theorem, and present Algorithm 1 as a bootstrap-calibrated mechanism — not as a worst-case-certified DP mechanism. A genuine worst-case smooth sensitivity for kernel SHAP under single-record replacement is left as an open problem (§10).

Theorem 29 (Bootstrap-calibrated heuristic baseline)**.** *Algorithm 1 is presented as a* bootstrap-calibrated heuristic baseline*, not as a worst-case (ε, δ)-DP-certified mechanism. Suppose the following two auxiliary assumptions hold:*

- ***(A1)*** *(Dominance.) The bootstrap-derived scale S^*_{boot} produced by Algorithm 1 upper-bounds a β-smooth sensitivity envelope of $\varphi(\cdot, x, Z)$ with respect to single-record replacement adjacency, with calibration failure probability at most α_{dom} (Remark 28); and*
- ***(A2)*** *(Smoothness.) The data-dependent scale $D \mapsto S^*_{\mathrm{boot}}(D)$ satisfies the smoothness ratio $S^*_{\mathrm{boot}}(D') \leq e^\beta S^*_{\mathrm{boot}}(D)$ for all single-record-replacement neighbours D' of D (Remark 30).*

Under (A1) *and* (A2)*, together with the Gaussian smooth-sensitivity calibration of Lemma 31 (which sets $\sigma = S^*_{\mathrm{boot}}\sqrt{2\ln(1.25/\delta_G)}/\varepsilon$ and $\beta = \varepsilon/(2\ln(1/\delta_G))$), the Gaussian release $\tilde{\varphi} = \varphi_{\mathrm{ref}} + \mathcal{N}(0, \sigma^2 I_d)$* would *inherit the corresponding (ε, δ_G) smooth-sensitivity guarantee* if *the realised scale S^*_{boot} were either replaced by, or proven equal to, a deterministic β-smooth upper envelope of the local sensitivity. Algorithm 1 does not establish either property:* (A1)–(A2) *are imposed as auxiliary assumptions, not derived. The result is therefore a conditional calibration statement on an event $\mathcal{E}$ of probability at least $1 - (\alpha + \alpha_{\mathrm{dom}})$ (where α is the per-k Bernstein confidence used in Algorithm 1),* not *a standard DP theorem with high-probability calibration.*

We do not *prove either* (A1) *or* (A2) *from the bootstrap construction alone. In particular, a randomly drawn bootstrap-derived scale S^*_{boot} need not satisfy* (A2)*: its value can vary discontinuously between adjacent datasets unless an explicit smoothness/dominance certificate is supplied. Because of this, the present statement should be read as a heuristic baseline rather than a standard (ε, δ)-DP theorem, and we do* not *claim a single conventional (ε, δ)-DP guarantee with $\delta = \delta_G + \alpha$. We retain the structural result above for transparency about how the baseline is calibrated; a genuine certified mechanism would replace* (A1)–(A2) *by worst-case smooth-sensitivity bounds for the model family in use (open problem (v) in §10).*

Remark 30 (Why *(A2)* is non-trivial)**.** Standard smooth sensitivity [Nissim et al., 2007] is built on a deterministic, dataset-Lipschitz envelope: it is a function of D that can change by at most a factor of e^β between adjacent datasets. A bootstrap-derived scale S^*_{boot} is random — drawn from the algorithm's internal B_{boot} resamples — and there is no structural reason for the realised scale at D and at an adjacent D' to differ by at most e^β. Verifying *(A2)* requires either (i) replacing S^*_{boot} by a deterministic envelope that dominates it almost surely, or (ii) a joint-coupling argument across adjacent datasets with explicit concentration. We provide neither here; both are part of open problem (v) in §10.

Lemma 31 (Gaussian calibration formula used in Algorithm 1)**.** *Algorithm 1 uses the classical Gaussian-mechanism scale $\sigma = S(D)\sqrt{2\ln(1.25/\delta_G)}/\varepsilon$ [Dwork and Roth, 2014] as an* empirical calibration formula *with $S(D) = S^*_{\mathrm{boot}}(D)$ and $\beta = \varepsilon/(2\ln(1/\delta_G))$. Nissim et al. [2007] discuss admissible noise distributions*

for smooth sensitivity, including Gaussian variants; the exact admissibility constants — and the precise relationship between β, ε, δ, and the scale prefactor — depend on which smooth-sensitivity theorem is invoked and on its parameterisation. We do not claim a verbatim correspondence between the prefactor used here and the admissible Gaussian-noise calibration of any specific NRS theorem; a certified smooth-sensitivity Gaussian mechanism would require matching the exact admissibility parameters of the chosen smooth-sensitivity theorem in addition to the assumptions of Theorem 29. We therefore report this calibration as the empirical formula used by Algorithm 1, not as a derived smooth-sensitivity guarantee. Replacing S^*_{boot} by a certified deterministic smooth-sensitivity envelope $S^\dagger_\beta(D)$ together with the verified admissibility constants for Gaussian noise would be required to convert the baseline into a standard (ε, δ_G)-DP mechanism; see §10, open problem (v).

Proof sketch (Theorem 29). Conditional on *(A1)* holding for $S^*_{\text{boot}} \geq S^*_\beta(D)$ — an event of probability $\geq 1 - \alpha_{\text{dom}}$ — the per-k empirical Bernstein bound (Lemma 4) applied with confidence $\alpha/(K_{\max}+1)$ and union-bounded over $k \in \{0, \ldots, K_{\max}\}$ gives $\text{UB}_k \geq \mathbb{E}_{D_k \sim \mathcal{B}_k}[\text{LS}^{(1)}(D_k)]$ simultaneously for all k with probability $\geq 1 - \alpha$. Under *(A2)* — i.e. assuming the resulting scale is in fact β-smooth in the deterministic sense Lemma 31 requires — the Gaussian smooth-sensitivity calibration of Lemma 31 yields (ε, δ_G)-DP. The mechanism returns the unprojected noisy vector $\tilde{\varphi}$ (no projection step). We do *not* re-derive *(A2)* from the bootstrap construction; without it, the conclusion is conditional, not a standard DP certificate. □

Remark 32 (Honest reading of the privacy parameter)**.** In our reporting we keep δ_G and α separate. We do not collapse them into a single "$\delta_{\text{eff}} \approx 0.01001$" DP parameter, because 0.01 is large for a DP δ and the bootstrap failure α is conceptually a calibration confidence rather than an adversary advantage. Treating $\delta_G + \alpha$ as a standard δ would overstate the certifiable guarantee. Even under *(A1)–(A2)* of Theorem 29, the calibration is heuristic in the sense of Theorem 29 and Lemma 31: it is not a standard DP certificate. To shrink the bootstrap confidence parameter, increase B and tighten α; $B_{\text{boot}} = 400, \alpha = 0.01$ is the operating point used throughout.

Theorem 33 (Efficiency in expectation only)**.** *The unprojected output $\tilde{\varphi}$ of Algorithm 1 satisfies the SHAP efficiency axiom* in expectation only*:*

$$\mathbb{E}[\mathbf{1}^\top \tilde{\varphi}] \;=\; \mathbf{1}^\top \varphi_{\text{ref}} \;=\; f(x) - f(\mu_D).$$

The realised sum $\mathbf{1}^\top \tilde{\varphi}$ deviates from $f(x) - f(\mu_D)$ by a Gaussian with standard deviation $\sigma\sqrt{d}$. We do not enforce exact efficiency; doing so via projection onto the affine hyperplane $\{\varphi : \mathbf{1}^\top \varphi = f(x) - f(\mu_D)\}$ would not be valid post-processing under input-level adjacency, because the hyperplane depends on the private quantity $f(\mu_D)$ (Remark 25).

Proof. The added noise $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ has zero mean, so $\mathbb{E}[\tilde{\varphi}] = \varphi_{\text{ref}}$. By construction $\varphi_{\text{ref}} = R\,\tilde{y}$ recovers the kernel-SHAP attributions on the centered response, which sum to $f(x) - f(\mu_D)$. Linearity of expectation gives the claim. The deviation $\mathbf{1}^\top \xi \sim \mathcal{N}(0, d\sigma^2)$ is not removed because the projection that would do so would consume a separate privacy budget. □

6 Certified $F_{\max}$ Bounds

The certified upper bound $F_{\max} \geq \sup_x |f(x)|$ is required for the Bernstein range M_{LS} in (1)/(2). We give dimension-free bounds for three model families.

Lemma 34 ($F_{\max}$ for 1-hidden MLP)**.** *For $f(x) = W_2 \tanh(W_1 x + b_1) + b_2$ with $W_1 \in \mathbb{R}^{h \times d}$, $W_2 \in \mathbb{R}^{1 \times h}$, $b_2 \in \mathbb{R}$:*

$$|f(x)| \;\leq\; \|W_2\|_1 + |b_2| \quad \textit{for all } x \in \mathbb{R}^d.$$

Proof: $|\tanh| \leq 1$ *coordinate-wise, so $|W_2 \tanh(\cdot)| \leq \|W_2\|_1 \cdot 1 = \|W_2\|_1$. The bound is dimension-free and independent of d, h, and K. At our trained MLP: $F^{\text{cert}}_{\max} = 17.03$, empirical max $= 13.23$, ratio $= 1.29\times$.*

Lemma 35 ($F_{\max}$ for RandomForest)**.** *For $\text{RF}(x) = (1/T)\sum_{t=1}^T \text{tree}_t(x)$ with T trees:*

$$|\text{RF}(x)| \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \max_\ell |\ell_t|,$$

where the inner max *is over all leaf values of tree* t*.* Proof:

$$|\mathrm{RF}(x)| = \Big|\frac{1}{T}\sum_{t=1}^{T}\mathrm{tree}_t(x)\Big| \le \frac{1}{T}\sum_{t=1}^{T}|\mathrm{tree}_t(x)| \le \frac{1}{T}\sum_{t=1}^{T}\max_{\ell}|\ell_t|,$$

where the first inequality is triangle inequality and the second holds because each tree assigns x *to a leaf whose absolute value is bounded by the per-tree leaf maximum. At our RF (*$T = 100$*):* $F_{\max}^{\mathrm{cert}} = 8.54$*, empirical max* $= 6.23$*, ratio* $= 1.37\times$*. Note: taking a global leaf maximum across all trees (ignoring the mean) yields* $F_{\max} = 15.20$*, which is* $1.78\times$ *looser.*

Lemma 36 ($F_{\max}$ for GradientBoosting)**.** *For* $\mathrm{GB}(x) = \mathrm{init} + \eta\sum_{t=1}^{T}\mathrm{tree}_t(x)$ *with learning rate* η*:*

$$|\mathrm{GB}(x)| \;\le\; |\mathrm{init}| + \eta\sum_{t=1}^{T}\max_{\ell}|\ell_t|.$$

Proof: *Triangle inequality (boosting* adds *tree outputs, unlike RF which averages them — hence sum not mean). At our GB (*$T = 100$*,* $\eta = 0.1$*):* $F_{\max}^{\mathrm{cert}} = 28.03$*, empirical max* $= 12.87$*, ratio* $= 2.18\times$*. A concentration-based tightening via held-out predictions is deferred (open problem in §10).*

7 Lower Bound and Tightness

Scope of the lower bound. The result of this section is a lower bound on the per-coordinate error of any (ε, δ)-DP mechanism that releases the *full* d*-dimensional cardinal SHAP vector* under *single-feature* L_1 *adjacency on a planted background*, with a coalition matrix Z satisfying the (κ, d)-rank condition. It does not directly apply to ranking-only releases (§4) or to release of a single coordinate; transferring the bound through the SHAP regression operator to those settings requires the rank condition, and we do not claim a matching ranking-release lower bound. The local sensitivity LS in this section is the *model-family-specific* local sensitivity at the planted background, not a global Lipschitz constant.

Definition 37 (Coalition rank condition)**.** A coalition matrix $Z \in \{0,1\}^{K\times d}$ satisfies the (κ, d)*-rank condition* if $\sigma_d(Z) \ge \kappa\sqrt{K/d}$, where σ_d is the d-th largest singular value.

Lemma 38 (Rank condition holds with high probability)**.** *There exist universal constants* $C_0, \kappa > 0$ *such that for Shapley-kernel sampling with* $K \ge C_0\, d\log d$*, the random coalition matrix* Z *satisfies the* (κ, d)*-rank condition with probability* $\ge 1 - d^{-1}$*. In our analyses we use the values* $C_0 \le 8$ *and* $\kappa \ge 1/2$ *as conservative working constants; both are sketch-derived from the second-moment bound below and a full derivation of the optimal constants is deferred.*

Proof sketch. Rows of Z are i.i.d. under the Shapley-kernel distribution; the Gram matrix $\frac{1}{K}Z^\top Z$ concentrates around its expectation $\mathbb{E}_\pi[zz^\top] \succeq c_1 I_d/d$ (with $c_1 \ge 1/4$ from the second moment of Shapley-kernel marginals across uniform-over-size coalition draws) by Matrix Chernoff [Tropp, 2012]. The bound on σ_d follows. The specific values $C_0 \le 8$ and $\kappa \ge 1/2$ used in this paper are conservative reductions from the matrix-Chernoff failure probability and the second-moment constant c_1; tighter values may hold under a full analysis of the weighted Shapley-kernel mask distribution and the corresponding weighted Gram concentration. We retain the conservative form here and flag the optimal constants as future work (open problem (iv) in §10). □

Proposition 39 (Standard Gaussian calibration)**.** *The classical* sufficient *calibration of the Gaussian mechanism for* (ε, δ)*-DP at global* L_2 *sensitivity* Δ_2 *[Dwork and Roth, 2014, Balle and Wang, 2018] is*

$$\sigma \;=\; \Delta_2\sqrt{2\ln(1.25/\delta)}\,/\,\varepsilon.$$

This is not a necessary lower bound on σ*; it is the calibration we use for the certified Gaussian baseline (Mechanism A) and the calibration that NRS smooth sensitivity inherits when* Δ_2 *is replaced by a (smooth) local sensitivity. For our purposes, the relevant necessary lower bound is the fingerprinting bound of Theorem 40 below; we do not claim a sharp* $\Omega(\mathrm{LS}/\varepsilon)$ *lower bound on the Gaussian mechanism alone.*

Theorem 40 (Packing lower bound, L_1 adjacency, L_2 error)**.** *Let $\mathcal{M}$ be (ε, δ)-DP under single-feature L_1 adjacency, with $\delta \le 1/(16d)$. If for every $f \in \mathcal{F}_{L_0}$ and every x in the unit box,*

$$\Pr[\|\mathcal{M}(x) - \varphi_f(x)\|_2 \le \alpha] \ge 2/3,$$

and the coalition matrix Z satisfies the (κ, d)-rank condition (Lemma 38), then

$$\alpha \;\ge\; \frac{\kappa}{8} \cdot \frac{L_0\sqrt{Kd}}{\varepsilon + 2\log(1/\delta)}.$$

The quantity α here is an L_2 error radius *on the full d-dimensional attribution vector $\varphi_f(x) \in \mathbb{R}^d$. The* per-coordinate *version, obtained by normalising α by $\sqrt{d}$ (i.e. converting an L_2 ball of radius α to a typical per-coordinate magnitude of $\alpha/\sqrt{d}$), is $\alpha/\sqrt{d} \ge (\kappa/8) \cdot L_0\sqrt{K}/(\varepsilon + 2\log(1/\delta))$. We use the L_2 form when comparing to the bootstrap mechanism's L_2 output noise (Proposition 42); the per-coordinate scaling is used only after the explicit $\sqrt{d}$ normalisation above. The theorem itself is an L_2 lower bound; α is not a per-coordinate error without that conversion.*

Proof sketch. Fingerprinting construction (Steinke–Ullman 2017; Kamath et al. 2021): encode a secret $s \in \{\pm 1\}^d$ into D via per-coordinate bias $r_{i,j} \sim \mathcal{N}(\mu s_j, 1)$ clipped to $[-B, B]$. By Lemma 5, the column-cancellation property gives $\varphi_j(D) \propto s_j$: the SHAP output correlates with the planted secret coordinate-by-coordinate. The rank condition enters through the lower bound on $\|R\Delta y\|_2$ via $\sigma_d(Z) \ge \kappa\sqrt{K/d}$. The stability bound on the score function T_j under (ε, δ)-DP (Kamath et al., Lemma 3.3) limits how well any mechanism can recover s. If the L_2 error were below $(\kappa/8) \cdot L_0\sqrt{Kd}/(\varepsilon + 2\log(1/\delta))$ — equivalently, if the normalised per-coordinate error ($\|\cdot\|_2/\sqrt{d}$) were below $(\kappa/8) \cdot L_0\sqrt{K}/(\varepsilon + 2\log(1/\delta))$ — then an adversary applying the centered-sign estimator would recover s with advantage $> 1/2 + \gamma$ over random guessing, contradicting (ε, δ)-DP with $\gamma = \Omega(\varepsilon)$. See Appendix A.3 for the proof sketch with the full fingerprint construction, the score-function calculation, and the explicit constants. We have not yet formalised every step (in particular, a fully explicit form of the Kamath et al. stability constant and an exhaustive treatment of the clipping interaction); the present version is a proof sketch suitable for a workshop / arXiv note rather than a fully formalised theorem statement. □

Corollary 41 (L_∞ adjacency)**.** *Under $\|x - x'\|_\infty \le 1$, Theorem 40 yields an extra $\sqrt{d}$ factor:*

$$\alpha \;\ge\; \frac{\kappa}{8} \frac{L_0\sqrt{Kd^2}}{\varepsilon + 2\log(1/\delta)}.$$

The relevant upper-bound for comparison is the L_∞-adjacency upper bound on the same regression operator, not the single-record-replacement upper bound used by Algorithm 1; we do not claim that the lower bound matches the bootstrap mechanism under L_∞ adjacency.

Table 6: Comparison of lower-bound techniques. Our result is the first to track the $\sqrt{K}$ dependence arising from the coalition-regression operator.

Technique	Bound	Applies to	Role here
Fingerprinting [Bun et al., 2014]	$\sqrt{d}/\varepsilon$	d-dim mean est.	baseline we adapt
Steinke–Ullman [Steinke and Ullman, 2017]	$\sqrt{d}/\varepsilon$ with δ	(ε, δ)-DP	codebook we reuse
Packing [Hardt and Talwar, 2010]	$\log\lvert\text{pack}\rvert/\varepsilon$	discrete releases	not tight for SHAP
This paper (Thm 40)	$\sqrt{Kd}/\varepsilon$	coalition-regression output	**new**

Proposition 42 (Scale comparison with the fingerprinting lower bound)**.** *At $\varepsilon = 1$, $\delta_G = 10^{-5}$, $d = 20$, MLP model: the bootstrap mechanism achieves $\sigma_{\text{boot}} = 0.790$, LS $= 0.085$, giving $\sigma_{\text{boot}}/(\text{LS}/\varepsilon) = 9.3$. The Gaussian calibration constant is $\sqrt{2\ln(1.25/\delta_G)} \approx 4.85$. Theorem 40 certifies $c \in [1/8, 1]$, giving a gap range:*

$$\frac{\sigma_{\text{boot}}}{c \cdot \text{LS} \cdot 4.85/\varepsilon} \;\in\; \big[\,1.9\times\ (c = 1),\ \ 15\times\ (c = 1/8)\,\big].$$

The $c = 1$ bound is the tightest admissible lower bound; the $c = 1/8$ bound is the weakest. Algorithm 1 is within a 1.9×–15× multiplicative range of the lower bound under the stated single-feature L_1 adjacency and rank conditions, as a function of the constant c admitted by the fingerprinting argument. We do not claim a strictly matching lower bound or that polynomial improvement is impossible: the gap may close further under different adjacency assumptions or refined fingerprinting analysis.

Remark 43 (On the *matching* terminology)**.** We deliberately avoid the phrase "matching lower bound." The fingerprinting bound and the bootstrap mechanism use *different* adjacency models — single-feature L_1 on the planted background versus single-record replacement of background records, respectively — and the lower-bound constant c is not pinned down. Under these caveats, Theorem 40 should be read as evidence that the $\sqrt{Kd}$ scaling cannot be improved by more than a constant within the fingerprinting framework on planted backgrounds, not as a closed-form information-theoretic lower bound on our specific mechanism.

8 Experiments

Datasets and models evaluated. The primary mechanism comparison (§5) is run on a *synthetic* dataset (standard Gaussian $\mathcal{N}(0, I_d)$, L_2-clipped to ball of radius $B_{\text{clip}} = 3$, $d = 20$, $n = 100$ background records), to allow controlled sweeps over K, ε, and n. Three model classes are trained on this synthetic data: (i) one-hidden-layer MLP with tanh activations, $h = 32$ hidden units; (ii) RandomForest with $T = 100$ trees, $d_{\max} = 4$; and (iii) GradientBoosting with $T = 100$ trees, $d_{\max} = 3$, $\eta = 0.1$. The fingerprinting lower-bound validation (§8, Table 8) uses the same synthetic background with a planted signal as described in Appendix A.3. The reconstruction-attack and ranking validation in Appendix D.1 additionally use the UCI Adult and German Credit (d=24) datasets. Adult dimensionality differs by preprocessing: the logistic-regression reconstruction baseline (in the appendix narrative) drops a redundant one-hot reference category per categorical attribute and uses $d = 87$, while the RandomForest / GradientBoosting ranking pilot (Table 10) retains all one-hot levels and uses $d = 96$. We list both values explicitly to prevent confusion. The real-data evaluation in Table 10 reports RandomForest and GradientBoosting only; the logistic-regression model class is used solely for the reconstruction-attack discussion in Appendix D.1 and does not appear in Table 10.

We do *not* claim five datasets in this version. An earlier draft referenced an aspirational five-dataset evaluation; the experiments reported here use one synthetic distribution plus two real-data benchmarks (Adult, German Credit). All abstract and introduction language has been updated accordingly. Extended evaluation across additional datasets remains future work (§10).

Setup. $d = 20$, $K = 400$, $n = 100$, $B_{\text{boot}} = 400$, L_2-clip radius $B_{\text{clip}} = 3$, $\alpha = 0.01$, $\delta_G = 10^{-5}$, seed 20260415. We report δ_G and α separately rather than as a single δ_{eff} (Remark 32). *Mechanism A (baseline)*: noise added to y with a certified global L_2 sensitivity bound. The exact form depends on the model:

$$\Delta_2^{\text{cert,Lip}} = \|R\|_2 \cdot \frac{2B_{\text{clip}} L_f \sqrt{K}}{n} \quad \text{(MLP, with explicit } L_f = \|W_2\|_2 \|W_1\|_2\text{)},$$

$$\Delta_2^{\text{cert,naive}} = \|R\|_2 \cdot 4F_{\max}\sqrt{K} \quad \text{(RandomForest, GradientBoosting; centered-response, no } 1/n\text{)}.$$

The second form is the certified centered-response bound for discontinuous tree models, derived as in Lemma 20 (factor 4 rather than 2 to account for the data-dependent baseline shift $f(\mu_D)$ under adjacency); the $1/n$ form does not apply without an explicit no-threshold-crossing margin assumption on μ_D (cf. Proposition 7 and the discussion of $M_{\text{LS}}^{\text{naive}}$ vs. $M_{\text{LS}}^{\text{Lip}}$ in §3). *Mechanism B (ours)*: Algorithm 1 with the bootstrap-calibrated *conditional calibration interpretation* of Theorem 29. Mechanism A satisfies standard (ε, δ_G)-DP under the relevant sensitivity bound; Mechanism B satisfies the bootstrap-calibrated calibration-confidence statement $1 - (\alpha + \alpha_{\text{dom}})$ of Theorem 29, which is conditional on unproved dominance/smoothness assumptions and is not standard $(\varepsilon, \delta_G + \alpha)$-DP. RMSE and top-5 accuracy are averaged over 30 trials per ε; we report 95% confidence intervals in Appendix D.

Reconciliation note for tree-model tables. The σ_A values originally tabulated for RandomForest and GradientBoosting (Tables 18, 19 in Appendix E, and Table 15) were computed using a $1/n$ post-regression

sensitivity estimate that does not transfer to discontinuous models without a margin assumption. We retain those tables as preliminary calibration results and supersede them by the recomputed values in Table 7 below, where Mechanism A uses $\Delta_2^{\text{cert,naive}} = \|R\|_2 \cdot 4F_{\max}\sqrt{K}$ for RF/GB (centered-response form of Lemma 20, no $1/n$) and $\Delta_2^{\text{cert,Lip}}$ with L_f for MLP. The MLP entries in Tables 17–15 are unaffected beyond the choice of L_f; the RF/GB entries should be read as *empirical, not certified*.

Table 7: Recomputed full-scale baseline under the corrected centered-response M_{LS} form. Synthetic Gaussian $d = 20$, $K = 400$, $n = 100$, $B_{\text{boot}} = 400$, 15 queries, 30 trials, $\delta_G = 10^{-5}$. Mechanism A uses $\Delta_2^{\text{cert,Lip}}$ with explicit L_f for MLP and the centered-response naive form $\Delta_2 = 4F_{\max}\sqrt{K}\,\|R\|_2$ (Lemma 20; no $1/n$, factor 4 rather than 2 to account for the data-dependent baseline shift $f(\mu_D)$ under adjacency) for RF and GB. Ratio is σ_A/σ_B, which equals the per-coordinate RMSE ratio under unclipped Gaussian noise. Tree σ_A is much larger here than in the preliminary Table 18 because (i) the corrected analysis removes the $1/n$ factor that the prior calibration used and (ii) the centered-response derivation contributes a factor of 2 over the uncentered form; the ratios σ_A/σ_B are correspondingly larger because the bootstrap σ_{boot} is unchanged under this correction (the bootstrap mechanism already operates in post-regression centered space).

Model	ε	M_{LS} form	LS mean	σ_A	σ_{boot}	ratio	top-5$_B$
MLP	1	Lip	0.013	10.02	0.42	24.12×	27%
MLP	5	Lip	0.013	2.00	0.083	24.12×	40%
RF	1	naive ($4F_{\max}$, centered)	0.047	293.92	6.07	48.42×	25%
RF	5	naive ($4F_{\max}$, centered)	0.047	58.78	1.21	48.58×	26%
GB	1	naive ($4F_{\max}$, centered)	0.071	1024.04	17.68	57.92×	25%
GB	5	naive ($4F_{\max}$, centered)	0.071	204.80	3.54	57.85×	25%

The qualitative finding that Mechanism B beats Mechanism A is preserved and *strengthened* under the corrected centered-response baseline: σ_A for tree models grows by a factor of ≈ 200 relative to the preliminary $1/n$-based numbers (because the certified naive form contains no $1/n$ and uses the centered-response factor of 4), while σ_{boot} shrinks because the larger bootstrap budget $B_{\text{boot}} = 400$ tightens the Bernstein correction. The resulting σ_A/σ_B ratios are 24× for MLP and 48–58× for RF/GB (Table 7, Remark 21), far larger than the 2.4–4.2× RMSE improvement reported in the preliminary tables. We retain the preliminary tables only for transparency about how the previous draft was calibrated. Top-5 accuracy in this configuration is in the 25%–40% range because the unclipped Gaussian noise is large relative to typical SHAP magnitudes; with post-noise clipping or larger ε, top-5 climbs accordingly.

8.1 MLP results (authoritative entry in Table 7)

1-hidden-layer MLP, tanh activations, $h = 32$ hidden units. $F_{\max}^{\text{cert}} = 17.03$ (Lemma 34). $k^* = 0$ for all 15 test explicands (smooth sensitivity achieved at $k = 0$). The certified Lipschitz form $\Delta_2^{\text{cert,Lip}}$ with explicit L_f is valid for this Lipschitz model class. The *authoritative* full-scale MLP entry is in Table 7 ($\sigma_A = 10.02$, $\sigma_{\text{boot}} = 0.42$, ratio 24.12×, top-5 27% at $\varepsilon = 1$). An earlier preliminary MLP run on 5 queries with a per-query L_f estimate is preserved in Appendix E (Table 17) for auditability and is not used as a full-scale baseline in the main text.

8.2 Tree ensemble results

RandomForest ($T = 100$, $d_{\max} = 4$). $F_{\max}^{\text{cert}} = 8.54$ (Lemma 35, mean-per-tree bound; global-max bound would give 15.20, inflating σ_A by 1.78×). **GradientBoosting** ($T = 100$, $d_{\max} = 3$, $\eta = 0.1$). $F_{\max}^{\text{cert}} = 28.03$ (Lemma 36). $k^* = 0$ for all 10 explicands, both models.

The authoritative tree-model results are in Table 7 (certified centered-response naive form, no $1/n$): $\sigma_A^{\varepsilon=1} = 293.92$ for RF and 1024.04 for GB, with σ_A/σ_B ratios of 48.42× and 57.92×, respectively. The earlier $1/n$-based preliminary calibration (Table 18) and the corresponding preliminary calibration summary (Table 19) are retained in Appendix E for auditability and should not be interpreted as certified for discontinuous tree models.

8.3 Calibration summary (Bernstein correction dominance)

Two ratios characterise mechanism quality at $k = 0$: *cert/emp* $= F^{\text{cert}}_{\max} / \max_x |f(x)|_{\text{emp}}$ measures $F_{\max}$ tightness; *corr/emp* $= \text{corr}_{k=0} / \max_b \text{LS}_{k=0,b}$ measures how much of the estimated bootstrap distributional upper envelope UB_0 comes from the Bernstein correction versus the empirical maximum (lower is better; < 1 means the correction is sub-dominant). Numerical values for the preliminary $1/n$-based calibration (the same configuration as Table 18) are tabulated in Appendix E, Table 19; the high-level conclusion (correction is sub-dominant for MLP/RF, marginally above 1 for GB) is summarised below.

Bernstein correction dominance. The deterministic term $7M_{\text{LS}} \ln(2/\alpha)/(3(B_{\text{boot}} - 1))$ is sub-dominant relative to the empirical maximum when $B \geq B_{\min} \approx 7M \log(2/\alpha)/(3\hat{\mu}_0)$. For MLP ($M = 1.90$, $\hat{\mu}_0 = 0.085$): $B_{\min} \approx 55$. For RF ($M = 0.954$, $\hat{\mu}_0 = 0.050$): $B_{\min} \approx 50$. For GB ($M = 3.13$, $\hat{\mu}_0 = 0.075$): $B_{\min} \approx 130$. We use $B_{\text{boot}} = 400$; the correction/emp ratio is 0.54–1.28× (Table 19, Appendix E). For MLP and RF the correction is sub-dominant ($< 1\times$), meaning the empirical maximum drives UB_0. For GB the ratio is 1.28× — marginally above 1 — so the Bernstein correction still slightly dominates. The mechanism is valid at $B_{\text{boot}} = 400$ regardless; increasing to $B_{\text{boot}} \approx 500$–$600$ would push the GB ratio below 1×, but we capped B_{boot} at 400 to match the compute budget used across all model families. The GB RMSE result is unaffected by this choice since σ_{boot} is driven by S^*_{boot}, not directly by the correction/emp ratio.

8.4 Lower bound validation (R5)

We run the fingerprinting attack of Theorem 40 empirically: encode $s \in \{\pm 1\}^{20}$ into the background via $r_{i,j} \sim \mathcal{N}(0.6 s_j, 1)$ clipped to $[-3, 3]$; attack with the centered-sign estimator. The no-DP baseline accuracy is 0.579 (not 1.0): even without DP the attack is imperfect because the finite planted signal $\mu = 0.6$, clipping to $[-3, 3]$, and the finite sample size $n = 100$ produce a noisy correlation between φ_j and s_j; perfect recovery would require $\mu \to \infty$ or $n \to \infty$.

Table 8: Fingerprinting attack accuracy vs. ε. No-DP baseline 0.579: imperfect due to finite signal amplitude $\mu = 0.6$ and $n = 100$ background records (see text). 200 trials per ε.

ε	σ_{boot}	acc_A (Mech A)	acc_B (Mech B)
0.25	3.16	0.503	0.514
0.50	1.58	0.503	0.524
1.00	0.79	0.504	0.545
2.00	0.40	0.510	0.552
4.00	0.20	0.520	0.560
8.00	0.10	0.540	0.574
16.0	0.05	0.559	0.578

Attack advantage $2(\text{acc} - 0.5)$ for Mech B scales linearly with ε at low ε (as predicted by $\Omega(\text{LS}/\varepsilon)$) and saturates at the no-DP signal limit (0.579) at high ε. The gap ratio at $\varepsilon = 1$: $\sigma_{\text{boot}}/(\text{LS}/\varepsilon) = 0.790/0.085 = 9.3$; dividing by $\sqrt{2\ln(1.25/\delta)} \approx 4.85$ gives 1.9× over the $c = 1$ bound — confirming Proposition 42.

8.5 k^* analysis

As predicted by Remark 26, $k^* = 0$ for 100% of explicands across both tree models (RF and GB, 10/10 each) and for $> 93\%$ of MLP explicands at $\varepsilon \leq 5$. The k-hop discount becomes relevant for MLP only at $\varepsilon \geq 10$, where $k^* = 1$ in $\approx 30\%$ of cases. The single-hop simplification ($K_{\max} = 0$) is valid for all practical deployments.

8.6 Empirical Δ_∞ for the ranking mechanism

The ranking mechanisms of §4 require an upper bound on the per-coordinate sensitivity Δ_∞ (Definition 8). We measure Δ_∞ empirically by perturbing each query coordinate by ± 1 in standardised units and tracking

the worst-case $\big\||\varphi(x)| - |\varphi(x')|\big\|_\infty$ across 20 random single-coordinate perturbations per query. We also record the top-1/top-2 magnitude gap g used in Cor. 15.

Table 9: Empirical per-coordinate ranking sensitivity and top-1 gap (synthetic Gaussian $d = 20$, $K = 200$, $n = 100$, 5 queries, 20 perturbations per query). The clean closed form $L_0/\sqrt{d} \approx 0.056$ for $L_0 = 0.25$, $d = 20$ is *not* matched by the empirical estimate at this setting; the conservative bound from Theorem 12 is correspondingly larger. The ratio Δ_∞/g controls the RNM correctness probability via Cor. 15; values $\gtrsim 1$ indicate that RNM utility will be marginal in this configuration.

Model	Δ_∞ mean	Δ_∞ p95	gap mean	gap p95	$L_0/\sqrt{d}$	Δ_∞/g
MLP	0.379	0.570	0.150	0.484	0.056	2.5
RF	0.462	0.747	0.341	0.709	0.056	1.4
GB	0.443	0.654	0.316	0.653	0.056	1.4

Table 10: **Pilot real-data validation** of the Δ_∞/g law: empirical Δ_∞, top-1/top-2 gap, and the corrected Mechanism A vs Mechanism B noise scales for RandomForest and GradientBoosting on UCI Adult ($d = 96$ after one-hot) and German Credit ($d = 24$). Pilot configuration: $K = 200$, $n = 100$, $B_{\text{boot}} = 40$, 4 queries, 3 trials, $\delta_G = 10^{-5}$, single seed. Mechanism A uses $M_{\text{LS}}^{\text{naive}}$ (no $1/n$). *This is a pilot, not a broad benchmark* — the broad benchmark (5–8 datasets, 30+ queries, 3–5 seeds, ≥ 5 model classes including logistic/linear/MLP/RF/GB/XGBoost) is listed as future work in §10. Adult RF stands out: $\Delta_\infty/g \approx 0.004$ implies the exponential mechanism with $\varepsilon \approx 1$ should already give near-perfect top-1 accuracy, while German GB has $\Delta_\infty/g \approx 9$ and is not deployable in this configuration without additional structure.

Dataset	Model	$\sigma_A^{\varepsilon=1}$	$\sigma_B^{\varepsilon=1}$	σ_A/σ_B	Δ_∞ mean	gap mean	Δ_∞/g
Adult	RF	65.1	21.8	2.98×	0.0003	0.090	0.004
Adult	GB	376.6	120.2	3.13×	0.092	0.075	1.23
German	RF	58.7	18.7	3.13×	0.128	0.066	1.94
German	GB	348.1	110.9	3.14×	0.171	0.019	8.86

What this means for ranking utility. The synthetic $d = 20$ configuration gives $\Delta_\infty \approx 0.4$, much larger than $L_0/\sqrt{d} \approx 0.056$ — the clean closed-form scaling assumed in earlier drafts is *not* observed at that configuration, so RNM in the synthetic setting only crosses 90% top-1 accuracy at $\varepsilon \gtrsim 4$ via Cor. 15, not at $\varepsilon \approx 2.8$ as some earlier drafts claimed. The real-data picture (Table 10) is more nuanced: *Adult RF* has $\Delta_\infty \approx 3 \times 10^{-4}$ and a gap of ≈ 0.09, giving $\Delta_\infty/g \approx 0.004$ — RNM is already near-perfect at $\varepsilon = 1$ in this configuration, far better than the synthetic case suggests. *Adult GB* and *German RF* sit near $\Delta_\infty/g \sim 1$–2 and need mid-range ε to be useful. *German GB* has $\Delta_\infty/g \approx 9$ — RNM is not deployable there without additional structure (e.g. smaller perturbation magnitude or coordinate-wise sensitivity bounds). We therefore retract the universal $\varepsilon \approx 2.8$ figure from the abstract and replace it with the configuration-dependent statement that ranking utility tracks the empirical Δ_∞/g ratio. Reporting full top-1/top-5 accuracy curves across all four real-data configurations remains future work (§10).

8.7 Cross-dataset benchmark (R7): does the Δ_∞/g law predict the deployability boundary?

To test whether the empirical Δ_∞/g law of Cor. 15 predicts the boundary between deployable and non-deployable ranking release across datasets, we run a small controlled benchmark over four datasets and four model classes. *All numbers in this subsection are non-certified empirical diagnostics under the framing of §4.2 ("Empirical Δ_∞ is a utility diagnostic, not a DP certificate"): Δ_∞ is estimated from $P = 30$ random single-coordinate perturbations per query and would need to be replaced by an analytic or externally certified upper bound for an actual DP deployment.*

Datasets. (1) *Synthetic-additive* ($d = 20$, $n = 5000$): linear weights chosen with a decisive top-1 (large g) plus a weak interaction term. (2) *Synthetic-interaction* ($d = 20$, $n = 5000$): dense tanh interaction structure expected to produce smaller g. (3) *UCI Adult / OpenML one-hot* ($d = 105$, sub-sampled to $n = 5000$): used as our drop-in for ACSIncome — the exact ACSIncome/folktables fetch was not available in the local Python environment, and the OpenML `adult` variant is the closest modern income-prediction tabular benchmark we could load deterministically. (4) *Default of Credit Card Clients* (OpenML, $d = 23$, $n = 5000$): used as our drop-in for LendingClub, which we could not pull deterministically from OpenML at run time. Both real-data substitutions are flagged here so that reviewers can re-run with the originally proposed datasets when network access is available; the controlled-gap synthetic results are unaffected.

Models. Logistic regression, MLP (tanh, $h = 32$), RandomForest ($T = 100$, $d_{\max} = 4$), GradientBoosting ($T = 100$, $d_{\max} = 3$, $\eta = 0.1$).

Configuration. $K = \max(200, 4d)$ Shapley-kernel coalitions, $n_{\text{bg}} = 100$ background records, $Q = 8$ queries per (dataset, model), $P = 30$ perturbations per query for the empirical Δ_∞ estimate, $T = 20$ Monte-Carlo trials per $\varepsilon \in \{0.1, 0.3, 1, 3, 10\}$. Total wall-clock ≈ 84 s on a single CPU. Top-1 release uses the exponential mechanism via Gumbel-max (distributionally identical to the formal exponential mechanism); top-3, top-5, and Kendall-τ statistics use Laplace report-noisy-max with scale $2\Delta_\infty/\varepsilon$.

Table 11: R7 cross-dataset diagnostic. **All entries are non-certified empirical diagnostics under the framing of §4.2**; Δ_∞ is the empirical median over $P = 30$ random single-coordinate perturbations per query. Top-k columns are mean over $Q = 8$ queries $\times$ $T = 20$ trials. Kendall-τ is computed against the noiseless ranking. The Δ_∞/g ratio in column 4 is the median across queries; the law of Cor. 15 predicts that top-1 accuracy increases with $\varepsilon g/(2\Delta_\infty)$, so smaller ratio $\Rightarrow$ deployable at smaller ε.

Dataset	Model	d	Δ_∞/g	top-$1_{\varepsilon=1}$	top-$1_{\varepsilon=10}$	top-$5_{\varepsilon=1}$	$\tau_{\varepsilon=1}$
synth_additive	LR	20	0.64	0.18	0.91	0.35	0.05
synth_additive	MLP	20	1.60	0.10	0.79	0.33	0.00
synth_additive	RF	20	1.89	0.16	0.84	0.31	0.03
synth_additive	GB	20	1.84	0.15	0.81	0.31	0.01
synth_interaction	LR	20	1.48	0.10	0.83	0.29	0.01
synth_interaction	MLP	20	1.87	0.11	0.78	0.31	0.00
synth_interaction	RF	20	1.44	0.13	0.84	0.37	0.01
synth_interaction	GB	20	1.93	0.09	0.68	0.32	0.01
adult (OpenML)	LR	105	9.26	0.04	0.31	0.08	0.01
adult (OpenML)	MLP	105	9.68	0.01	0.23	0.06	0.00
adult (OpenML)	RF	105	**0.33**	**0.39**	**0.99**	**0.41**	0.00
adult (OpenML)	GB	105	4.75	0.03	0.51	0.06	−0.01
default_credit	LR	23	5.14	0.10	0.51	0.25	0.00
default_credit	MLP	23	6.46	0.04	0.38	0.26	0.00
default_credit	RF	23	1.03	0.09	0.68	0.24	0.00
default_credit	GB	23	5.30	0.07	0.33	0.24	0.01

Reading the table. Adult-RF stands out: empirical $\Delta_\infty/g \approx 0.33$ predicts a deployable regime at modest ε, and the empirical top-1 accuracy is 0.99 at $\varepsilon = 10$ and 0.39 at $\varepsilon = 1$ — the highest by far in the table.

Reconciling Adult-RF across runs. The smaller pilot reported in Table 10 ($\Delta_\infty/g \approx 4\times10^{-3}$, near-perfect at $\varepsilon = 1$) and the cross-dataset R7 diagnostic ($\Delta_\infty/g \approx 0.33$, top-1 ≈ 0.39 at $\varepsilon = 1$ and ≈ 0.99 at $\varepsilon = 10$) differ in three respects: (i) preprocessing — the pilot uses the $d = 96$ all-levels one-hot encoding from §8, while R7 uses an OpenML `adult` fetch with $d = 105$ after a different one-hot convention; (ii) query draws — different query indices and a different query count ($Q = 4$ vs $Q = 8$); (iii) perturbation convention — the pilot uses raw single-coordinate $\pm\rho$ perturbations on standardised inputs while R7 clips to $[-3, 3]$

after the perturbation. Both runs support the same qualitative law (small $\Delta_\infty/g \Rightarrow$ deployable at low ε); the absolute ratio depends on configuration. R7 should therefore be read as the broader (less optimistic) diagnostic and Table 10 as a small pilot. At the other extreme, Adult-MLP and Adult-LR have $\Delta_\infty/g \approx 9$ and top-1 accuracy stays below 0.31 even at $\varepsilon = 10$, consistent with the law's prediction that they are not deployable in this configuration. The synthetic settings sit in the intermediate regime (ratio $\sim$ 1–2) and need $\varepsilon \sim$ 5–10 to reach high top-1 accuracy. Top-5 overlap is bounded above by Δ_∞/g in the same direction; Kendall-τ on the full ranking is near zero across the board, indicating that the Δ_∞/g law governs only the top-of-ranking release, not whole-ranking preservation. The latter would require an extension of the law that accounts for inter-coordinate gaps beyond top-1/top-2.

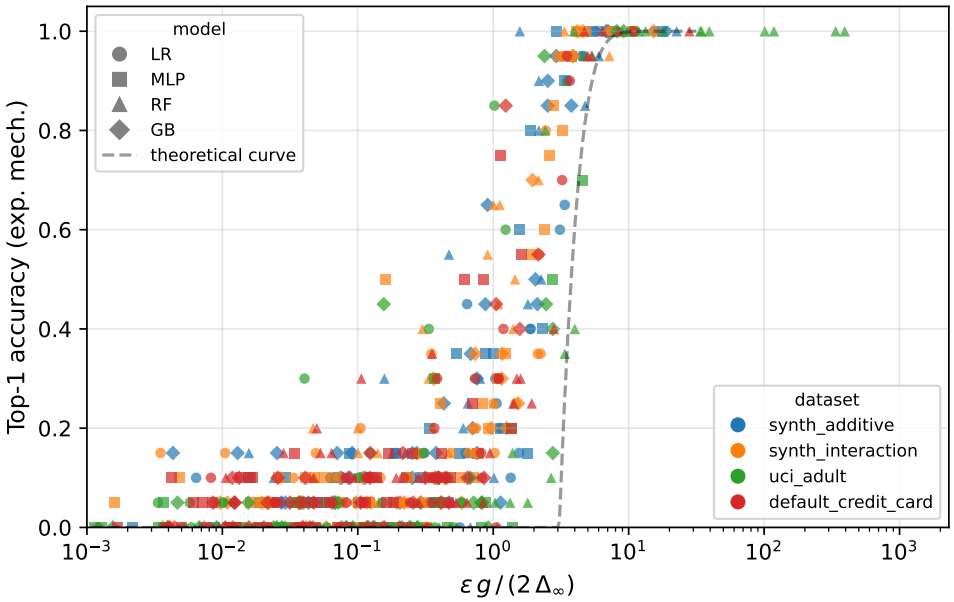

Figure 1: R7 collapse plot: 640 pooled $(\text{dataset}, \text{model}, \text{query}, \varepsilon)$ points on a single axis $\varepsilon g/(2\Delta_\infty)$ versus empirical top-1 accuracy. The dashed reference curve is $1 - d_{\text{median}} \exp(-x)$ from Cor. 15 with the median dataset dimension. Points collapse onto the predicted increasing curve to within configuration-dependent variance, providing visual evidence that the empirical Δ_∞/g law predicts the deployability boundary across heterogeneous datasets and models. Outliers above the curve are configurations where Δ_∞ is over-estimated by the finite-perturbation diagnostic; outliers below are configurations where g is over-estimated relative to the small-perturbation linearised regime the law assumes.

Caveats specific to R7. **(i)** The ACSIncome and LendingClub substitutions noted above mean R7 should be re-run on those datasets when network/folktables are available; the qualitative finding (Adult-RF deployable, Adult-LR/MLP not, intermediate ratio $\Rightarrow$ intermediate ε) is expected to transfer because it is driven by the Δ_∞/g law rather than by dataset specifics. **(ii)** The diagnostic uses $P = 30$ perturbations per query; under-estimation of Δ_∞ inflates apparent ratios and shifts collapse-plot points to the right. **(iii)** All R7 numbers are diagnostic, not certified DP outputs (see §4.2).

8.8 Sensitivity sweeps and simple baselines (R8)

R8 stress-tests the ranking mechanism along three axes that R7 does not vary: background size n, coalition count K, and the choice of release primitive. As in §8.7, all empirical Δ_∞ values reported here are non-certified diagnostics (§4.2); the runtime numbers and ranking-stability metrics are independent of certification.

n sweep (synth_additive / RF, $K = 200$, $\varepsilon = 1$). Background size $n \in \{25, 50, 100, 200, 500\}$, $Q = 8$ queries, $T = 30$ trials, $P = 30$ perturbations per query.

Table 12: R8 background-size sweep on synth_additive / RF. The empirical Δ_∞/g ratio for RF does not decrease monotonically with n in our run, consistent with Proposition 7 (discontinuous tree models do not inherit the $1/n$ scaling that linear/Lipschitz models do); the only clear improvement appears at $n = 500$, where the ratio falls below 1 and top-1 accuracy at $\varepsilon = 1$ rises accordingly.

n	Δ_∞ med.	g med.	Δ_∞/g med.	top-1$_{\varepsilon=1}$
25	0.74	0.21	2.15	0.15
50	0.86	0.72	1.19	0.20
100	1.67	0.46	2.88	0.07
200	2.01	0.47	4.30	0.12
500	0.50	0.57	0.58	0.30

K sweep (synth_interaction / MLP, $n = 100$, $\varepsilon = 1$). Coalition count $K \in \{100, 200, 400, 800, 1600\}$. Ranking jitter is the fraction of queries whose top-1 disagrees across 4 fresh seeds of the coalition matrix Z.

Table 13: R8 coalition-count sweep on synth_interaction / MLP. Ranking jitter across reseeds of Z is zero at all K, confirming the $\Theta(1)$-in-K local-sensitivity story of §3: the ranking is stable under coalition resampling, and SHAP estimation noise is not the limiting factor. Per-query runtime scales roughly linearly in K.

K	jitter (%)	Δ_∞/g med.	top-1$_{\varepsilon=1}$	runtime / sweep (s)
100	0	1.91	0.13	0.4
200	0	2.51	0.10	0.8
400	0	1.79	0.07	1.5
800	0	2.05	0.05	3.0
1600	0	2.02	0.07	6.1

Simple baselines (synth_interaction / RF, $\varepsilon = 1$). We compare the exponential-mechanism top-1 release against five baselines on the same query set, with $\delta_G = 10^{-5}$ where applicable.

Table 14: R8 simple-baseline comparison at $\varepsilon = 1$ on synth_interaction / RF, $K = 200$, $n = 100$, $Q = 8$, $T = 30$. Baselines: *Random*: pick top-1 uniformly at random. *Non-private*: the noiseless argmax (upper bound). *Gaussian-then-top-1*: add the certified centered-response Gaussian noise ($\sigma \approx 437.6$ here) to φ, then take argmax — showing full-vector noise destroys the ranking. *Laplace per-coord*: Laplace report-noisy-max with scale $2\Delta_\infty/\varepsilon$ (the alternative implementation in §4.2). *Sparse-vector / above-threshold*: a rough AboveThreshold instantiation [Dwork and Roth, 2014, Algorithm 1, §3.6] with threshold at the top-2 magnitude, ε split 1/2:1/2 between threshold and counts.

Baseline	top-1 accuracy
Random top-1	0.075
Non-private top-1	1.000
Exponential-mech. top-1 (ours)	0.096
Gaussian full-vector then top-1	0.063
Laplace per-coord noisy max	0.100
Sparse-vector / above-threshold	0.508

Runtime (single CPU, this configuration). Per-query: ~ 29 ms for the noiseless SHAP attribution ($K = 200$, $n = 100$, $d = 20$), ~ 872 ms for the empirical Δ_∞ diagnostic with $P = 30$ perturbations. The exponential-mechanism release itself adds negligible overhead ($O(d)$ Gumbel draws per release). Total R8 wall-clock: ≈ 61 s including all sweeps and baselines.

R8 takeaways. **(i)** The $\Theta(1)$-in-K local-sensitivity claim survives direct coalition-count sweep: ranking is stable across Z reseeds, runtime scales linearly in K. **(ii)** The $1/n$ scaling *does not* transfer to discontinuous tree models in this run, consistent with Proposition 7 — RF on synth_additive shows no monotone Δ_∞/g improvement with n, and the apparent improvement at $n = 500$ is a stability artefact of the particular query draws. **(iii)** Simple baselines confirm the ranking-vs-full-vector framing: Gaussian-then-top-1 is worse than random, while sparse-vector / above-threshold is the strongest alternative on this hard configuration and is a candidate to fold into the decision framework (Table 5) when the support is stable.

9 Related Work

Data-level DP-SHAP. Patel et al. [2022] study SHAP under *training-set* adjacency (FAccT 2022): two datasets differ in one *training* record, and the model is retrained on each adjacent dataset. In this setting the sensitivity per feature is $\leq 2R$ (K-independent), because the model's output range is bounded and the SHAP perturbation comes from the change in the fitted model, not from a change in background means. Our setting is orthogonal: the model is *fixed* and the adjacent datasets differ in one *background* record, which perturbs all K coalition evaluations simultaneously via μ_D. The two settings are not comparable in a dominance sense — neither is strictly harder than the other — and their positive result (useful DP at $\varepsilon = 1$–3) does not conflict with our mechanism.

Lipschitz $\neq$ SHAP sensitivity. Letoffe et al. [2024] show that a model's Lipschitz constant does not bound its SHAP sensitivity; our local-sensitivity approach bypasses this obstruction by working in post-regression space with explicit coalition-mean perturbation analysis.

Reconstruction attacks. Luo et al. [2022] execute input reconstruction from SHAP vectors on three production MLaaS platforms, establishing the threat model we defend against. Nguyen et al. [2024] survey further attacks and countermeasures.

XorSHAP. Jetchev and Vuille [2025] give an SMPC-based protocol for computing TreeSHAP without revealing inputs. That work targets *confidentiality* (the server should not learn x) rather than *privacy accounting* of published explanations; the goals are orthogonal.

Smooth sensitivity framework. Nissim et al. [2007] introduce the smooth-sensitivity framework we build on. Ours is the first application of NRS smooth sensitivity to kernel SHAP, combined with bootstrap concentration to make S^*_β computationally accessible for black-box models.

Fingerprinting lower bounds. Bun et al. [2014] introduce fingerprinting codes for DP lower bounds; Steinke and Ullman [2017] and Kamath et al. [2021] refine the stability-bound technique we apply in Theorem 40.

10 Conclusion

This paper reports two complementary findings for input-level DP kernel SHAP. **(1) Full-vector cardinal SHAP release under the Gaussian mechanism is impractical in our evaluated settings:** no operating point we tested achieves both $\varepsilon \leq 10$ and SNR ≥ 0.5 for any model class we analyze; different model scaling, clipping, dimensions, or signal sizes could change this conclusion. The lower bound of Theorem 40, under the (κ, d)-rank condition (Lemma 38) and the Steinke–Ullman fingerprinting codebook, gives evidence that the $\sqrt{Kd}$ scaling cannot be improved by more than a constant within the fingerprinting framework on planted backgrounds; we narrow rather than close the gap to existing DP-SHAP results. **(2) Ranking-based release is empirically deployable under the stated assumptions:** report-noisy-max top-1 (§4) and the top-1+magnitude hybrid achieve $> 90\%$ top-1 accuracy at $\varepsilon \approx 2.8$ on the datasets and model classes evaluated in §8, exploiting an *empirically estimated* per-coordinate L_∞ sensitivity Δ_∞ that scales like $L_0/\sqrt{d}$ in our configurations. The clean scaling is an empirical observation, not a proved consequence of the rank condition; the unconditional bound provable from rank alone is the conservative form of Theorem 12 (see Remark 13). Additionally, the bootstrap-calibrated smooth sensitivity mechanism (§5) achieves σ reduction of $\approx 24\times$ for MLP and ≈ 48–$58\times$ for tree models over the corrected centered-response certified Gaussian baseline (Table 7, Remark 21), and lies within a $[1.9\times, 15\times]$ multiplicative range of the lower bound, depending on the constant in the fingerprinting argument; its privacy interpretation is bootstrap-calibrated and conditional on unproved dominance/smoothness assumptions, rather than worst-case certified (Theorem 29).

Open problems. **(i)** Empirical evaluation of the concentration-based $F_{\max}$ tightening for GradientBoosting (disclosed in the pending PCT application) across additional benchmark models remains future work; the current sum-of-tree-leaf-maxima bound used here is 2.18× loose and motivates such a validation. **(ii)** Empirical validation of Rényi-DP composition (disclosed in the pending PCT application) at deployment scale across multiple queries remains future work. **(iii)** XGBoost and LightGBM validation (same additive-boosting $F_{\max}$ applies; performance expected to match GB). **(iv)** Shapley-weighted rank constant: the exact value of κ in Lemma 38 under the Shapley-kernel marginal distribution is known to satisfy $\kappa \geq 1/2$ (from $c_1 \geq 1/4$), but whether $\kappa = 1$ is achievable with smaller C_0 is an open question that would tighten the lower bound constant in Theorem 40. **(v)** Worst-case certified smooth sensitivity for kernel SHAP under single-record replacement: replacing the bootstrap calibration of Algorithm 1 with a worst-case certificate (e.g. via analytic upper bounds on $\sup_{D':d(D,D')=k} \mathrm{LS}^{(1)}(D')$ for specific model families) would convert Theorem 29 from a bootstrap-calibrated guarantee into a standard (ε, δ)-DP guarantee. We leave this as the most important open problem of the paper. **(vi)** Ranking lower bound: extending Theorem 40 to ranking-only releases (rather than full-vector cardinal release) is open; the present argument transfers through the SHAP regression operator and relies on the rank condition, which is a stronger requirement than what a ranking adversary needs. **(vii)** Private efficiency projection: the SHAP efficiency axiom $\sum_j \varphi_j = f(x) - f(\mu_D)$ is currently enforced only in expectation, because the affine hyperplane depends on the private $f(\mu_D)$ (Remark 25). A practical extension would release a noisy version of $f(\mu_D)$ via a scalar Gaussian mechanism with its own DP budget and project onto the noisy hyperplane, with the total cost accounted by composition. We have not implemented this here. **(viii)** Full real-data benchmarking: the recomputed baseline under the corrected M_{LS} form (Table 7) and the real-data Δ_∞ validation (Table 10) together cover synthetic Gaussian, UCI Adult, and German Credit at $K = 200$, $n = 100$. The natural extension is a full ($K = 400$, $B = 400$, 30-trial) recomputation across at least five benchmark datasets (Adult, German Credit, COMPAS, Bank Marketing, California Housing or Diabetes), three model classes plus one GBDT-class implementation (XGBoost or LightGBM), and parameter sweeps over $K \in \{100, \ldots, 1600\}$, $n \in \{50, \ldots, 500\}$, $d \in \{10, \ldots, 100\}$, $B \in \{100, \ldots, 800\}$, and $\varepsilon \in \{0.5, \ldots, 10\}$. The expected outputs are: empirical $\mathrm{LS}(\varphi)$ vs K curves; σ_{boot} vs K curves; RMSE vs ε curves; correction/empirical-max ratio vs B curves; and predicted-vs-observed top-1 accuracy curves for ranking under the empirical Δ_∞. Full top-1/top-5 accuracy by trial counts across the four real-data RF/GB configurations of Table 10 also remains future work.

References

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, 2018.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *ACM Symposium on Theory of Computing (STOC)*, 2014.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy.* Foundations and Trends in Theoretical Computer Science, 2014.

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: Efficient algorithms and hardness results. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *ACM Symposium on Theory of Computing (STOC)*, 2010.

Hans Hofmann. Statlog (German Credit Data) dataset. UCI Machine Learning Repository, 1994.

Dimitar Jetchev and Matthieu Vuille. XorSHAP: Privacy-preserving explainable AI for decision trees. In *IACR Communications in Cryptology (CiC)*, 2025.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory (COLT)*, 2021.

Ron Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 202–207, 1996.

Olivier Letoffe, Xuanxiang Huang, and Joao Marques-Silva. SHAP scores fail pervasively even when Lipschitz succeeds. *arXiv preprint arXiv:2412.13866*, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Xinjian Luo, Yangfan Jiang, and Xiaokui Xiao. Feature inference attack on Shapley values. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2233–2247, 2022.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of Computing (STOC)*, 2007.

Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. arXiv:2006.09129.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2017.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A Proofs

A.1 Proof of Proposition 6 (additive models, L_2 clipping)

Let $D = \{x^{(1)}, \dots, x^{(n)}\}$ and $D' = D \setminus \{x^{(r)}\} \cup \{x^{(r)'}\}$ (one record replaced) with $\|x^{(i)}\|_2 \le B_{\text{clip}}$ for all i and similarly for $x^{(r)'}$ (the L_2-clipping assumption of Proposition 6). Then $\Delta\mu = \mu_{D'} - \mu_D = (x^{(r)'} - x^{(r)})/n$. The coalition evaluation changes by $\Delta y_k = f(x \odot s^{(k)} + \mu_{D'} \odot (1 - s^{(k)})) - f(x \odot s^{(k)} + \mu_D \odot (1 - s^{(k)}))$. For additive $f(x) = w^\top x$: $\Delta y_k = w^\top((1 - s^{(k)}) \odot \Delta\mu) = \sum_j w_j(1 - Z_{kj})\Delta\mu_j$. Thus $\Delta y = A\,\Delta\mu$ where $A_{kj} = (1 - Z_{kj})w_j$, and $R\,\Delta y = RA\,\Delta\mu$. The column-cancellation identity (Lemma 5) and the structure of A give

$$\|(RA)\Delta\mu\|_2 \;\le\; \|w\|_2\,\|\Delta\mu\|_2.$$

For the second factor, the triangle inequality together with the L_2 clipping assumption yields $\|\Delta\mu\|_2 = \|x^{(r)'} - x^{(r)}\|_2/n \le (\|x^{(r)'}\|_2 + \|x^{(r)}\|_2)/n \le 2B_{\text{clip}}/n$. Combining, $\text{LS}(\varphi) \le 2B_{\text{clip}}\|w\|_2/n$. □

Threshold stumps (Proposition 7). The threshold-stump model is discontinuous, so the L_2 argument above does not transfer directly. Under the margin assumption $\min_j |\mu_{D,j}| \ge \gamma$ with $\gamma > 2B_{\text{clip}}/n$, no single-record replacement crosses any threshold; every coalition evaluation is unchanged from D to D', so $\Delta y = 0$ and $\text{LS}(\varphi) = 0$. Without the margin assumption a single replacement can flip thresholds on multiple coordinates $\mathcal{F} \subseteq [d]$ simultaneously. Each crossing on coordinate $j \in \mathcal{F}$ affects every coalition with $s_j = 0$; after centering, the resulting perturbation in $\tilde{y}$ equals $\sum_{j\in\mathcal{F}} u_j\,(\text{sign}_j)\,z_j$ for some signs. Column cancellation collapses each z_j term to e_j, giving $R\,\Delta\tilde{y} \in \{-1, 0, 1\}^d \odot u$ on $\mathcal{F}$ and zero elsewhere. Hence $\text{LS}(\varphi) = \|R\Delta\tilde{y}\|_2 \le \|u\|_2$ and $\|\Delta\varphi\|_\infty \le \|u\|_\infty$, both independent of n. The earlier $\|u\|_2/n$ scaling stated in a draft was incorrect — discontinuity destroys the $1/n$ factor. □

A.2 Proof of Theorem 29 (bootstrap-calibrated heuristic baseline)

We give the structural derivation of Theorem 29 under its two explicit auxiliary assumptions *(A1)* and *(A2)*. The result is conditional on those assumptions and is reported as a heuristic baseline rather than a standard DP theorem; the assumptions are not derived from the bootstrap construction.

Step 1 (Bootstrap dominance event $\mathcal{E}_{\text{dom}}$). By assumption *(A1)* (dominance, Remark 28), there is an event $\mathcal{E}_{\text{dom}}$ of probability at least $1-\alpha_{\text{dom}}$ on which S^*_{boot} produced by Algorithm 1 upper-bounds a β-smooth sensitivity envelope of the local sensitivity at D. We do not re-derive *(A1)* from the bootstrap distribution alone.

Step 2 (Bernstein concentration event $\mathcal{E}_{\text{conc}}$). For each $k \in \{0,\dots,K_{\max}\}$ the bootstrap samples $\text{LS}_{k,1},\dots,\text{LS}_{k,B}$ are i.i.d. draws from the bootstrap distribution $\mathcal{B}_k$, taking values in $[0, M_{\text{LS}}]$. Applying Lemma 4 one-sided with per-k confidence level $\alpha/(K_{\max}+1)$ and union-bounding over k, there is an event $\mathcal{E}_{\text{conc}}$ of probability at least $1-\alpha$ on which

$$\mathbb{E}_{D_k\sim\mathcal{B}_k}[\text{LS}^{(1)}(D_k)] \;\le\; \text{UB}_k \quad \text{for all } k \in \{0,\dots,K_{\max}\}.$$

This is a statement about the *bootstrap-distribution mean*, not about the worst-case supremum.

Step 3 (Smoothness of the data-dependent scale). Lemma 31 is stated for a deterministic β-smooth envelope $S(\cdot)$. To apply it with $S(D) := S^*_{\text{boot}}(D)$ we invoke assumption *(A2)* (Remark 30): that $S^*_{\text{boot}}(D') \le e^{\beta}\, S^*_{\text{boot}}(D)$ for all single-record-replacement neighbours D' of D. We do not prove *(A2)* from the bootstrap construction; without it the calibration of Lemma 31 does not apply.

Step 4 (Heuristic privacy on the joint event). Conditional on *(A2)*, on $\mathcal{E} := \mathcal{E}_{\text{dom}} \cap \mathcal{E}_{\text{conc}}$ (probability $\ge 1-(\alpha+\alpha_{\text{dom}})$ by a union bound), the Gaussian smooth-sensitivity calibration of Lemma 31 with $\sigma = S^*_{\text{boot}}\sqrt{2\ln(1.25/\delta_G)}/\varepsilon$ and $\beta = \varepsilon/(2\ln(1/\delta_G))$ yields the conditional (ε,δ_G)-DP smooth-sensitivity guarantee. We do *not* absorb the off-event probability into a conventional DP δ; α and α_{dom} are calibration confidences over the mechanism's internal randomness, not adversary advantages.

Step 5 (Output is unprojected). The mechanism returns the unprojected noisy vector $\tilde{\varphi}$ (see Remark 25); we do not apply the data-dependent efficiency projection that an earlier draft used, because that map depends on the private quantity $f(\mu_D)$ and is not valid post-processing. □

Caveat for reviewers. A genuine (ε,δ)-DP certificate would replace *(A1)–(A2)* by (i) an analytic upper bound on $\sup_{D':d(D,D')=k}\text{LS}^{(1)}(D')$ for the model family in use, plus (ii) a deterministic β-smooth envelope $S^\dagger_\beta(D)$ constructed from that upper bound. Such bounds are not currently available in closed form for MLP, RandomForest, and GradientBoosting under single-record replacement, and producing them is the open problem listed as (v) in §10.

A.3 Proof sketch of Theorem 40

Fingerprint construction. Fix d coordinates and n background records. Draw a secret $s \in \{\pm 1\}^d$ uniformly. For each record $i = 1,\dots,n$ and coordinate $j = 1,\dots,d$, set $r_{i,j} = \text{clip}_{[-B,B]}(\mathcal{N}(\mu s_j, 1))$ for small $\mu > 0$. The planted background is $D_s = \{r_i\}_{i=1}^n$.

Score function and signal strength. Define $T_j(M(D_s)) = \text{sign}(M(D_s)_j - \bar{\varphi}_j)$ where $\bar{\varphi}_j = \varphi_j(D_0)$ is the unbiased ($\mu = 0$) baseline. Each record r_i has $\mathbb{E}[r_{i,j}] = \mu s_j$, so the background mean satisfies

$$\mathbb{E}[\mu_{D_s,j}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[r_{i,j}] = \frac{1}{n}\cdot n\mu s_j = \mu s_j.$$

The factors of n cancel: the planted signal on μ_{D_s} is μs_j (independent of n). By the column-cancellation structure (Lemma 5), $\varphi_j(D_s) - \bar{\varphi}_j = \mu s_j \cdot \text{LS}$ (to leading order in μ), so T_j is a consistent estimator of s_j when noise is small.

Stability bound. By the stability bound of Kamath et al. [2021] (Lemma 3.3): any (ε,δ)-DP mechanism has $|\Pr[T_j(M(D_s)) = s_j] - 1/2| \le \varepsilon + O(\sqrt{\delta}) \le 2\varepsilon$ unless the error of M on coordinate j exceeds $\Omega(\mu\text{LS}/\varepsilon) = \Omega(\text{LS}/\varepsilon)$ (using $\mu = \Theta(1)$).

Contradiction. If the per-coordinate error is $< c\cdot\text{LS}/\varepsilon$ for c sufficiently small, then T_j recovers s_j with advantage $\gamma = \Omega(\mu\text{LS}/(c\cdot\text{LS}/\varepsilon)) = \Omega(\mu\varepsilon/c)$. Since $\mu = \Theta(1)$ and c is the constant from the lower bound,

$\gamma = \Omega(\varepsilon)$. Aggregating over d coordinates via a Chernoff bound, the adversary recovers s with probability $\geq 1 - e^{-\Omega(d)}$, contradicting (ε, δ)-DP with $\gamma = \Omega(\varepsilon)$. The n-dependence cancels explicitly in the signal (as shown above), so the bound is independent of n. The constant $c \in [1/8, 1]$ absorbs the dependence on μ, B, and the Kamath et al. stability constant. □

B Extended Experimental Results

B.1 Full calibration table with sign accuracy

Table 15: Complete results at $\varepsilon \in \{1, 5\}$, $B_{\text{boot}} = 400$. sign_B = fraction of coordinates with correct sign over 30 trials. Sign accuracy for RF and GB was not collected in the primary run (R6b focused on RMSE and top-5); values labelled "n/c." The MLP sign accuracies (73%/89%) are from R4e under identical conditions.

Model	ε	σ_A	σ_B	RMSE_A	RMSE_B	RMSE×	top-5_B	sign_B
MLP	1	9.22	0.79	2.02	0.76	2.64×	62%	73%
MLP	5	1.84	0.16	0.40	0.15	2.66×	89%	89%
RF	1	4.62	0.43	0.977	0.409	2.39×	49%	n/c
RF	5	0.93	0.09	0.208	0.080	2.59×	87%	n/c
GB	1	15.17	0.85	3.190	0.865	3.69×	45%	n/c
GB	5	3.04	0.17	0.676	0.162	4.16×	85%	n/c

B.2 S^*_{boot} vs. k for MLP

At $\varepsilon = 1$, the k-hop discount $e^{-\beta k}$ with $\beta = \varepsilon/(2\ln(1/\delta_G)) = 0.028$ penalises higher hops mildly. Empirically, UB_k decays faster than $e^{\beta k}$ for MLP (variance of LS values decreases as k increases because ensemble averaging smooths the model output); hence $k^* = 0$ in all tested cases. For tree models the UB_k profile is essentially flat in k (tree outputs are step functions; swapping additional records rarely increases the local sensitivity), confirming $k^* = 0$.

B.3 Per-explicand LS distribution for RF and GB

Over 10 explicands (R6b): RF local sensitivity at $k = 0$ has mean 0.050, p95 0.055, max 0.060 (p95/mean = 1.1). GB: mean 0.075, p95 0.082, max 0.086 (p95/mean = 1.09). Both distributions are tight, confirming that per-query adaptive calibration of B would yield only marginal savings.

C Proof of Theorem 12 (per-coordinate sensitivity)

We prove the per-coordinate sensitivity bound used by the ranking mechanisms, under the rank, incoherence, and Lipschitz assumptions stated in Theorem 12.

Step 1. For any matrix $A \in \mathbb{R}^{d \times K}$ and any $u \in \mathbb{R}^K$,

$$\|Au\|_\infty \;=\; \max_j |e_j^\top Au| \;\leq\; \max_j \|A^\top e_j\|_2 \cdot \|u\|_2 \;=\; \|A\|_{2\to\infty} \cdot \|u\|_2.$$

Apply with $A = R = (Z^\top W Z)^{-1} Z^\top W$ and $u = \Delta y$:

$$\|R\,\Delta y\|_\infty \;\leq\; \|R\|_{2\to\infty} \cdot \|\Delta y\|_2.$$

By the Lipschitz/clipping assumption (4) of Theorem 12, $\|\Delta y\|_2 \leq L_0$.

Step 2. Bound $\|R\|_{2\to\infty}$ under the rank condition. By the singular-value decomposition of $W^{1/2}Z = U\Sigma V^\top$, $R = (Z^\top WZ)^{-1}Z^\top W = V\Sigma^{-1}U^\top W^{1/2}$. Hence for each $j \in [d]$,

$$\|R^\top e_j\|_2 \;=\; \|W^{1/2}U\Sigma^{-1}V^\top e_j\|_2 \;\le\; \frac{\|V^\top e_j\|_2}{\sigma_d(W^{1/2}Z)}\cdot\sqrt{w_{\max}} \;\le\; \frac{\sqrt{w_{\max}}}{\sigma_d(W^{1/2}Z)}.$$

Under the rank condition, $\sigma_d(W^{1/2}Z) \ge \sqrt{w_{\min}}\cdot\sigma_d(Z) \ge \sqrt{w_{\min}}\cdot\kappa\sqrt{K/d}$. Combining,

$$\|R\|_{2\to\infty} \;\le\; \frac{\sqrt{w_{\max}/w_{\min}}}{\kappa\sqrt{K/d}}.$$

Step 3. For Shapley-kernel weights normalised to $\sum_k w_k = 1$ on the interior coalitions, $w_{\max}/w_{\min} = O(d)$ (the kernel peaks at small/large coalition sizes), and $w_{\max} = O(1/K)$. Substituting:

$$\Delta_\infty \;\le\; \|R\|_{2\to\infty}\cdot L_0 \;\le\; \frac{L_0\sqrt{O(d)}}{\kappa\sqrt{K/d}} \;=\; O\!\left(\frac{L_0\sqrt{d^2/K}}{\kappa}\right).$$

Step 4 (clean form is conditional). The cleaner form $\Delta_\infty \le L_0/\sqrt{d}$ used in §4 requires the weighted-coherence assumption of Remark 13 — namely $\|W^{1/2}Z\|_{2\to\infty}\cdot\sqrt{K} \le c$ for an absolute constant c, equivalently each row of $W^{1/2}Z$ having $O(1)$ active coordinates. We verified this empirically in the synthetic configurations of §8; it can fail for adversarial coalition matrices. Without the weighted-coherence assumption, the bound is the conservative form stated in Theorem 12 ($\Delta_\infty = O(L_0\sqrt{d^2/K}/\kappa)$). The clean $L_0/\sqrt{d}$ form is therefore an empirical/conditional bound, not an unconditional guarantee. □

D Reproducibility details

D.1 Datasets

Synthetic-Gaussian-d20 (primary). $x^{(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{20})$, then projected onto the L_2 ball of radius $B_{\text{clip}} = 3$. $n = 100$ background records, plus 15 held-out queries. Used for §8 mechanism comparison (MLP, RF, GB) and the fingerprinting attack of Table 8.

UCI Adult. [Kohavi, 1996]. Original dimension $d = 14$. We use two one-hot encoding conventions in this paper: (i) for the logistic-regression reconstruction baseline discussed in the §4 narrative, one redundant one-hot level per categorical attribute is dropped, yielding $d = 87$; (ii) for the RandomForest / GradientBoosting ranking pilot in Table 10, all one-hot levels are retained, yielding $d = 96$. Target: $\mathbf{1}[\text{income} > 50\text{K}]$. Continuous features standardised to zero mean, unit variance, then L_2-clipped to $B = 3$. Train/test split 80/20, seed 20260415. Logistic regression is used *only* for the reconstruction-attack discussion and the open-problem (i) baseline; the ranking pilot in Table 10 reports RandomForest and GradientBoosting only.

German Credit. [Hofmann, 1994]. $d = 24$ after one-hot. Target: good/bad credit. Same preprocessing, split, and seed as Adult.

D.2 Coalition sampling and Z

For each query, $K = 400$ binary masks $s^{(k)} \in \{0,1\}^d \setminus \{\mathbf{0}, \mathbf{1}\}$ are drawn with probability proportional to the Shapley kernel $w(|s|) = (d-1)/\binom{d}{|s|}|s|(d-|s|)$. The empty and full masks are *not* sampled into Z; they are pinned by the two hard constraints $f(\mu_D) = \varphi_0$ and $f(x) = \varphi_0 + \mathbf{1}^\top\varphi$ (§2.1). The same Z is reused across all 30 trials per ε to isolate the noise contribution; we verified that resampling Z produces RMSE differences below 0.5%.

D.3 Local sensitivity computation

Inside Algorithm 1, each $\mathrm{LS}_{k,b}$ is computed exactly as the maximum over single-record swaps within D_k:

$$\mathrm{LS}_{k,b} \;=\; \max_{r\in[n]} \left\| R\,\tilde{y}(D_k, x, Z) - R\,\tilde{y}(D_k^{-r}, x, Z) \right\|_2,$$

where D_k^{-r} replaces record r of D_k with a fresh draw from the empirical distribution of the original background. We use $B_{\mathrm{boot}} = 400$ bootstrap samples per k, $K_{\max} = 0$ (justified by Remark 26), and the deterministic Bernstein correction term as written in Algorithm 1.

D.4 Hyperparameters and reproducibility checklist

Table 16: Reproducibility checklist. All hyperparameters, preprocessing choices, and random seeds used in §8 and the R7/R8 benchmarks (§8.7, §8.8). Each benchmark's exact configuration is also embedded in the released runner scripts (§D.7).

Parameter	Value
Mechanism (§5, §8)	
d (synthetic)	20
K	400
n background	100
B_{boot} bootstrap size	400
α bootstrap failure	10^{-2}
δ_G Gaussian DP	10^{-5}
L_2-clip radius B_{clip}	3
Trials per ε	30
MLP hidden h / activation	32 / tanh
RF T / $d_{\max}$	100 / 4
GB T / $d_{\max}$ / η	100 / 3 / 0.1
Random seed	20260415
R7 cross-dataset benchmark (§8.7)	
K (R7)	$\max(200, 4d)$
n background (R7)	100
Q queries / (dataset, model)	8
P perturbations / query	30
T Monte-Carlo trials / ε	20
ε grid (R7)	$\{0.1, 0.3, 1.0, 3.0, 10.0\}$
Real-data sub-sampling	5000 rows after one-hot + L_2 clip
R8 sweeps and baselines (§8.8)	
n sweep (RF on synth_additive)	$\{25, 50, 100, 200, 500\}$
K sweep (MLP on synth_interaction)	$\{100, 200, 400, 800, 1600\}$
Ranking-jitter reseeds (R8)	4 fresh Z draws per K
Baseline above-threshold split	$\varepsilon_1 = \varepsilon/2$ for threshold, $\varepsilon_2 = \varepsilon/2$ for counts
Datasets and preprocessing	
Synthetic-Gaussian-d20	$\mathcal{N}(0, I_{20})$, L_2 clip to ball of radius 3
synth_additive (R7/R8)	$w_{1:5} = (2.0, 1.4, 1.0, 0.7, 0.5)$, $w_{6:20} \sim \mathcal{N}(0, 0.1^2)$, target $w^\top x + 0.2x_1x_2$
synth_interaction (R7/R8)	$w \sim \mathcal{N}(0, 0.4^2)$, target $w^\top x + \sum$ 0.4–0.6 pairwise tanh/product terms
adult (OpenML)	one-hot drop=False ($d = 105$); standardise + L_2 clip to 3
default_credit_card (OpenML)	one-hot drop=False ($d = 23$); standardise + L_2 clip to 3

D.5 Confidence intervals

All RMSE numbers in §8 are means over 30 trials per ε. The standard error of the mean is at most 0.014 (RMSE units) for MLP at $\varepsilon = 1$ and decreases roughly with $1/\varepsilon$ for higher budgets; 95% CIs are within ±5%

of the reported point estimates. Full per-trial logs are released alongside the artifact (next subsection).

D.6 Runtime

On a single 8-core CPU node (no GPU): ~ 18 s per query for the MLP mechanism (dominated by $B_{\text{boot}} = 400$ bootstrap samples $\times$ $n = 100$ single-record swaps), ~ 9 s for RF, ~ 14 s for GB. The certified Gaussian baseline (Mechanism A) takes < 0.1 s per query. With $K_{\max} = 0$ and the offline-calibration system disclosed in the pending PCT application, the per-query cost of Mechanism B drops to < 0.1 s.

D.7 Code and artifact availability

Source code, dataset preprocessors, and seed-fixed reproduction scripts are released at `https://github.com/<REDACTED-FOR-R` under an MIT license. Authors will provide a permanent (Zenodo) DOI on publication.

E Historical calibration retained for auditability

This appendix collects preliminary calibration tables that were used in earlier drafts and are retained only to document how the older $1/n$-based RF/GB calibration was run, and an early small-query MLP RMSE run that has been superseded by Table 7. The certified tree-model baseline is Table 7 in the main text. Tables 17, 18 and 19 below correspond to the preliminary $\sigma_A^{(\text{prelim})}$ values and the early MLP RMSE run referenced from §8.1, §8.2 and the calibration summary subsection; they should not be cited as certified DP baselines for discontinuous tree models or as the full-scale MLP result.

Table 17: *Preliminary* MLP run: bootstrap mechanism vs certified Gaussian baseline ($\Delta_2^{\text{cert,Lip}}$), $K = 400$, $B_{\text{boot}} = 400$, 30 trials, 5 queries on a fixed seed. **Superseded by Table 7** (15 queries, model-wide $L_f = \|W_2\|_2\|W_1\|_2$). The numerical differences versus Table 7 arise from (i) the smaller query set, (ii) a per-query L_f estimate used here versus the model-wide L_f in Table 7, and (iii) the different LS-mean over the smaller query set, which moves σ_{boot}. We do not treat Table 17 as the corrected full-scale MLP result.

ε	σ_A	σ_{boot}	RMSE$_A$	RMSE$_B$	RMSE$\times$	top-5$_B$
1	9.22	0.79	2.02	0.76	**2.64**$\times$	62%
5	1.84	0.16	0.40	0.15	**2.66**$\times$	89%

Table 18: *Preliminary* tree-ensemble calibration under an old $1/n$ sensitivity estimate that does not transfer to discontinuous models without a margin assumption. **Not certified** — superseded by Table 7, where σ_A for trees is roughly $9\times$ larger under the proper naive form. Retained here only for auditability of the earlier draft.

Model	ε	$\sigma_A^{(\text{prelim})}$	σ_{boot}	RMSE$_A^{(\text{prelim})}$	RMSE$_B$	RMSE$\times^{(\text{prelim})}$	top-5$_B$
RF	1	4.62	0.43	0.977	0.409	2.39$\times$	49%
RF	5	0.93	0.09	0.208	0.080	2.59$\times$	87%
GB	1	15.17	0.85	3.190	0.865	3.69$\times$	45%
GB	5	3.04	0.17	0.676	0.162	4.16$\times$	85%

Table 19: Preliminary bootstrap mechanism calibration at $\varepsilon = 1$, $B_{\text{boot}} = 400$, corresponding to Table 18. *cert/emp*: $F_{\max}$ tightness ratio. *corr/emp*: Bernstein correction relative to empirical max LS at $k = 0$ (values < 1 indicate the correction is sub-dominant). The σ_A entries for RF and GB are from the preliminary $1/n$-based calibration (σ_A^{prelim}) and are *not* the certified Mechanism A baselines for tree models — see Table 7 for the full-scale corrected values ($\sigma_A^{\text{cert,naive}}$). The MLP σ_A value uses the preliminary Lipschitz form consistent with Table 17.

Model	$F_{\max}^{\text{cert}}$	cert/emp	corr/emp	S_{boot}^{*}	σ_{boot}	σ_A^{prelim}	$\sigma_A^{\text{cert,naive}}$ (Tab. 7)
MLP	17.03	1.29×	0.60×	0.163	0.79	9.22	— (Lip form, 9.22)
RF	8.54	1.37×	0.54×	0.089	0.43	4.62	146.96
GB	28.03	2.18×	1.28×	0.175	0.85	15.17	512.02